

Gütekriterien der nichtmetrischen Multidimensionalen Skalierung

Inaugural-Dissertation
zur Erlangung des Grades eines
Doktors der Wirtschafts- und Sozialwissenschaften
bei der Wirtschafts- und Sozialwissenschaftlichen
Fakultät der Freien Universität Berlin

vorgelegt von
Dipl.math. Bernd Kolleck
aus Hannover

Berlin, 1982





Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41546618_0024

Gütekriterien der nichtmetrischen
Multidimensionalen Skalierung

Inaugural-Dissertation
zur Erlangung des Grades eines
Doktors der Wirtschafts- und Sozialwissenschaften
bei der Wirtschafts- und Sozialwissenschaftlichen
Fakultät der Freien Universität Berlin

vorgelegt von
Dipl.math. Bernd Kolleck
aus Hannover

Berlin, 1982

1. Gutachter: Prof. Dr. Horst Skarabis

2. Gutachter: Prof. Dr. Rolf Ebbighausen

Tag der mündlichen Prüfung: 28.1.1983

Gewidmet meinem Vater,

dem Dipl.-Agr. Horst Kolleck,

der diese Arbeit vielfältig unterstützt hat.

Zusammenfassung

Statistische Gütekriterien für komplexe Modelle wie die Multidimensionale Skalierung sind bisher nur unter einschneidenden Verteilungsannahmen möglich gewesen. Wir wollen hier zeigen, wie sich aus der ordinalen Struktur des Datenmaterials eine quasi "natürliche" Formulierung des statistischen Entscheidungsproblems ergibt. Daraus abgeleitet können Risikofunktionen entwickelt sowie gezeigt werden, wie gesichert Entscheidungen bei unterschiedlichen Streßwerten, Dimensionen und Objektzahlen (Punktezahlen) möglich sind. Für die Benutzung eines Maßes, das die Entfernung zwischen Daten und MDS-Skalierung auf ordinaler Basis mißt, lassen sich Gründe der Logik (von der Datenstruktur her), der Analogie (zu verwandten metrischen Maßen) und der statistischen Handbarkeit finden.

Zwar kann das hier beschriebene Vorgehen nicht die substanzwissenschaftlich notwendige Begründung für die Anwendung der Methode ersetzen, aber es bietet statistische Hilfsmittel für die Entscheidung dazu.

Inhaltsverzeichnis

Zusammenfassung	2
1. Anwendungsprobleme der Multidimensionalen Skalierung	5
2. Verfahren der nichtmetrischen MDS, Streßmaße	25
3. Entscheidung zwischen MDS-Konfigurationen	42
3.1 Hotelling's t	42
3.2 Intervallskalierte "Measures of Fit"	46
3.3 Ordinale Maße	48
3.4 Beziehungen zwischen metrischen und ordinalen Korrelationsmaßen	52
4. Hypothesentests auf Mengen von Rangfolgen	66
4.1 Testverfahren	66
4.2 Definitionen	69
4.3 Darstellung der Rangfolgen als Graph, Beispiele	71
4.4 Verteilungen	77
4.5 Statistische Formulierung des Testproblems	87
5. Verteilungen von Rangfolgen, Risikofunktionen	90
5.1 Verteilung der Distanzen zwischen darstellbaren Rangfolgen	90
5.2 Anzahl der Rangfolgen in Klassen mit gleichem Streß	100
5.3 Anzahl der Rangfolgen mit gleichem Abstand zu einer bestimmten Rangfolge	107

5.4 Verteilungsannahmen für Rangfolgen	110
5.5 Schätzung der Risiken bei Verwendung des Permutationsabstandes	114
5.6 Schätzung der Risiken bei der Verwen- dung von Kruskal's Streß	128
6. Weitere Ergebnisse der Monte-Carlo Studie	138
Anhang	144
Literaturliste	162

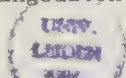
1. Kapitel

Anwendungsprobleme der Multidimensionalen Skalierung

Unter dem Begriff "Multidimensionale Skalierung" (MDS) werden Verfahren zusammengefaßt, die metrische oder ordinale Daten in räumliche Distanzen transformieren. Die Ausgangsdaten sind empirische Angaben zu Ähnlichkeiten oder Unähnlichkeiten zwischen Objekten. Diese Menge von Angaben ist mindestens ordinaler Natur und kann also als Rangfolge geschrieben werden. Skalierung bedeutet nun, zu jedem Objekt einen Punkt in einem Raum festgelegter Dimension p zu finden, so daß die metrischen Distanzen zwischen den Punkten den Unähnlichkeiten zwischen den Objekten entsprechen. Je größer die Unähnlichkeit zwischen zwei Objekten ist, desto größer soll die Distanz zwischen den sie repräsentierenden Punkten sein. Die Rangfolge der Distanzen im R^p soll der Rangfolge der Unähnlichkeiten möglichst genau entsprechen, die beiden Rangfolgen sollen sich also möglichst "isoton" zueinander verhalten.

Ein Maß für den Grad der Anpassung ist das sogenannte Streßmaß, das im nächsten Kapitel näher beschrieben und untersucht wird. Es ist eine Summe von normierten Quadraten von Differenzen ("Sum of Squares") zwischen den Distanzen einer optimalen Skalierung der Ausgangsdaten und - nicht etwa den Ausgangsdaten selbst, denn die können ja ordinal sein, sondern - metrischen Größen, die in ihrer Rangfolge genau den empirischen Daten entsprechen, zusätzlich aber den entsprechenden räumlichen Distanzen möglichst nahe kommen. Diese Größen sind die Schätzwerte einer isotonen Regression.

Dadurch ist erreicht, daß der Grad der Übereinstimmung von der Rangfolge der Ausgangsdaten und der Rangfolge



der Distanzen eine wesentliche Komponente ist, die sich im Streßmaß ausdrückt und vom Streßmaß bezeichnet wird. Eine weitere Eigenschaft, die das Streßmaß auszeichnet, ist seine metrische Natur, es mißt Differenzen zwischen metrischen Größen. Inwieweit diese Eigenschaft den möglicherweise ordinalen Ausgangsdaten gerecht wird, wurde bisher nicht untersucht.

Die sich in Anbetracht eines Skalierungs-Ergebnisses aufdrängende Frage, wie denn diese erstaunliche Transformation ordinaler in metrische Daten vonstatten gehen kann, wird in Kapitel 2 ausführlicher beantwortet. Hier sei nur darauf hingewiesen, daß die räumliche Konfiguration, das Ergebnis einer MDS, durch schrittweises Verschieben von Punkten hergestellt wird, die so lange bewegt werden, bis die Distanzen die Unähnlichkeiten gut wiedergeben. Als Optimalitätskriterium und "Richtungsweiser" dieser iterativen Operation dient wiederum das Streßmaß.

Bekannte Beispiele für gelungene Multidimensionale Skalierungen sind die Skalierung von Farbvergleichen bei Shepard (36) und die Skalierung von Fehlern beim Hören von Morsezeichen nach Shepard und Kruskal (19). Bei den Farbvergleichen wurden Individuen gebeten, die Ähnlichkeiten zwischen je zwei von 14 Farben zu benennen, also $\binom{14}{2} = 91$ Paarvergleiche durchzuführen (Ekman, (9)). Die Skalierung ergab den psychologischen Farbenkreis.

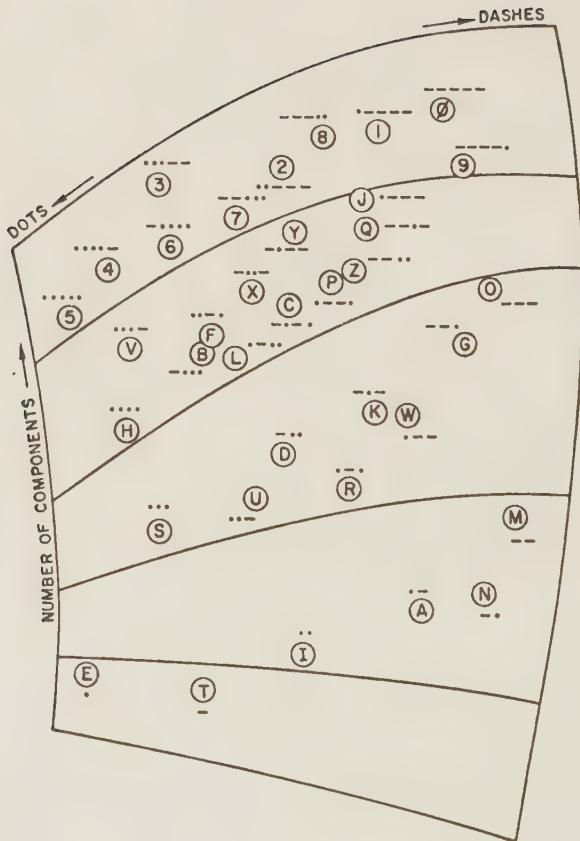


FIGURE 13

Spatial configuration for Ekman's 14 colors as reconstructed in the present reanalysis

aus: (36)

Die Skalierung von Morsecode-Zeichen zeigt Zeichen eng zusammenliegend, deren Verwechslungshäufigkeit groß ist. Hier wird die Vermutung zur Gewißheit, daß ähnliche Zeichen häufiger verwechselt werden als unähnliche. Die empirischen Daten stammen von Rotkopf (31).



Shepard's Interpretation of Morse Code Configuration

aus: (51)

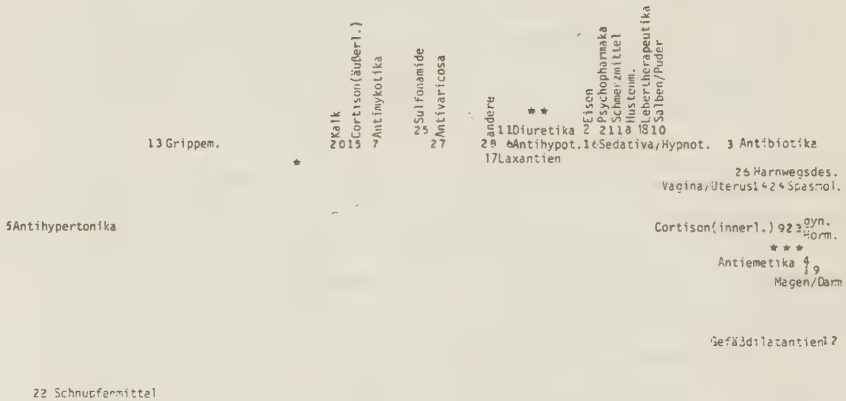
Beide Beispiele ließen Skalierungen im $\mathbb{R}^2$ zu, und das ist gut so, denn ansonsten wäre die graphische Wiedergabe schwierig.

Generell wird von einer Skalierung gewünscht, daß sie in Räumen niedriger Dimension, d.h. höchstens dreidi-

mensional, als äußerstes vierdimensional, erfolgt. Skalierung dient vor allem anderen der Sichtbarmachung des Datengehalts. Sie stellt weniger Anforderungen an mathematische Stringenz als an gefällige Repräsentation, sie ist weniger analytisch als darstellend. MDS-Verfahren kommen darin den Maximen der "Datenanalyse" gleich, daß sie die Daten eher selbst sprechen lassen als komplizierte Wahrscheinlichkeits- bzw. Signifikanzsätze über sie zu formulieren. Andererseits kann man die MDS sicher allein wegen ihres komplizierten Algorithmus nicht zur Methodensammlung der "Datenanalytiker" zählen.

Als Anwendungsbeispiel aus der Medizinsoziologie sei hier noch eine Skalierung von ärztlich verschriebenen Medikamenten bei gleichzeitig gestellten Diagnosen gezeigt. Ausgangsdaten sind eine Kontingenztafel mit 24 Medikamentengruppen und drei Diagnosen aus dem Bereich psychosomatisch zu interpretierender Erkrankungen. Die Daten wurden Abrechnungsscheinen für die ärztliche Betreuung schwangerer Frauen und eingelösten Rezepten entnommen. Aufgrund der Daten war keine Zuordnung von Medikamenten zu Diagnosen möglich. In einem gegebenen Zeitraum wurden häufig mehrere Diagnosen gestellt und verschiedene Medikamente verschrieben. Aber es war zu erwarten, daß sich über die Vielzahl der Betreuungen - es wurden 2465 Abrechnungsscheine erhoben - ein Zuordnungsraster ergeben würde. Einen ersten Einblick soll die Skalierung der Medikamente ergeben. Die Unähnlichkeiten sind in diesem Fall die Korrelationskoeffizienten zwischen je zwei Medikamenten, berechnet über die drei Diagnosen.

* SCHWANGERSCHAFTSBESCHWERDEN
 ** VEGETATIVE DYSTONIE
 *** HYPEREMESIS



Darstellung des Zusammenhangs der verordneten Medikamente mit den Diagnosen 'Schwangerschaftsbeschwerden', 'vegetative Dystonie' und 'Hyperemesis'

Der Streß der Konfiguration beträgt ca. 0,09. Aus dem kurvilinearen Zusammenhang geht hervor, daß es sich um eine im wesentlichen eindimensionale Konfiguration handelt. Eng beieinander liegen die Medikamentengruppen Antiemetika und gynäkologische Hormone, die häufig im Zusammenhang mit der Diagnose Hyperemesis verab-

reicht werden. Plausibel ist auch der enge Zusammenhang zwischen Psychopharmaka, Schmerzmittel und Antihypotonika, der über die Diagnose der Vegetativen Dystonie vermittelt ist. Die räumliche Nähe von Diuretika, Antivaricosa und unspezifischen weiteren Medikamenten unterstützt aber die Annahme eines wenig spezifischen Umgangs mit dieser Diagnose.

Medikamente, deren Zusammenhang durch die ebenfalls recht unspezifische Diagnose Schwangerschaftsbeschwerden begründet ist, liegen verstreut im linken Teil der Grafik: Antihypertonika, Kalk, Antivaricosa und Eisen.

Diese Beispiele erscheinen überzeugend und einleuchtend. Auf MDS-Darstellungen kann man in einem Blick Vieles von Zusammenhängen erfassen, das ansonsten vielleicht auch bei eingehender Betrachtung der Rohdaten nicht klar zum Vorschein käme. Trotzdem prägte Roger Shepard, dem der große Durchbruch beim algorithmischen Lösungsverfahren der MDS gelungen ist, den Satz: "Ich bin zu der Überzeugung gekommen, daß die meisten Entdeckungen, die bis jetzt gemacht wurden, im Prinzip ohne die Hilfe der Multidimensionalen Skalierung hätten gemacht werden können." (38).

Bevor aber weitere Kritik an der Methode formuliert wird, sollen zunächst einige der Autoren, die der MDS einen großen Stellenwert einräumen, zitiert werden.

Anwendungen in der Soziologie: Facettentheorie

Louis Guttman entwickelte in den späten fünfziger Jahren Vorstellungen von theoriegebundener empirischer Forschung, die für die MDS Bedeutung bekommen sollten:

Die Facettentheorie (12). Der Begriff der Facette steht hier für so etwas wie eine Kategorie, nach der der Forscher sein Erhebungsdesign plant. Beispielsweise wurde im Jahre 1976 von Levy (23) eine Studie zum Thema "Zufriedenheit mit den Lebensbedingungen" veröffentlicht, die Fragen zu den Design-Kategorien oder Facetten

A = Modalität mit den Ausprägungen "Zustand" und "Ressourcen", sowie

B = Lebensbereiche mit den Ausprägungen "Gesundheit", "Arbeit", "Wohnen", "Ausbildung" etc.

enthielt. Jede Frage an die Interviewten wurde so formuliert, daß sie eindeutig je einer Ausprägung von beiden Facetten zuzuordnen war. Beispielsweise entspricht etwa die Frage

"Sind Sie mit Ihrer jetzigen Arbeit

{
sehr zufrieden
.....
.....
.....
sehr unzufrieden"

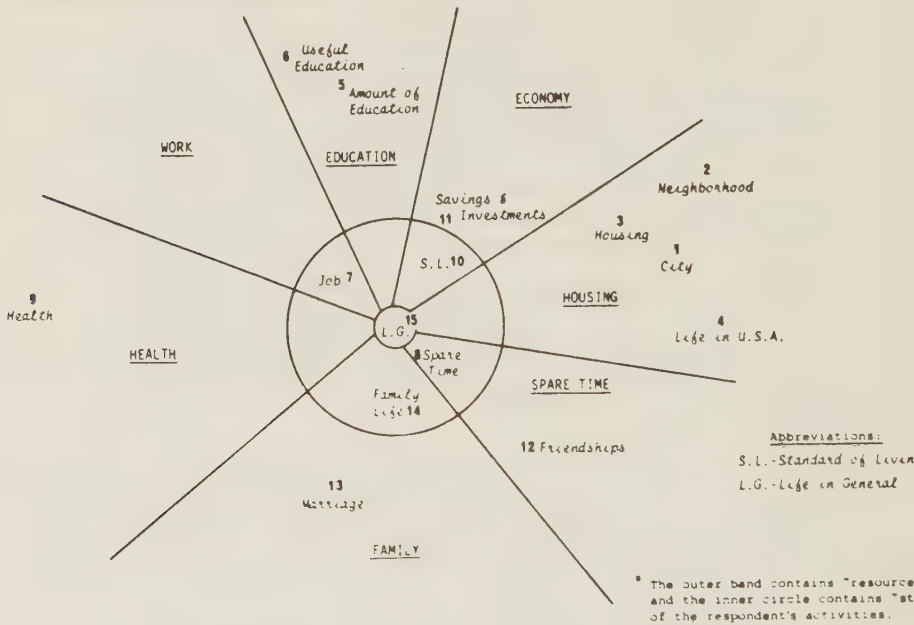
den Ausprägungen "Zustand" der Facette A und "Arbeit" der Facette B. Die Frage nach der Zufriedenheit mit dem Ausbildungsniveau wäre der Kategorie A "Resource" und B "Ausbildung" zuzuordnen.

Die Studie wurde in den USA mit 2.164 und in Israel mit 1.620 Interviews durchgeführt. Levy (23) erwartete, daß die von ihr geprägten Facetten für das Forschungsdesign im Antwortverhalten der Befragten wiederzufinden seien, d.h. daß die Datenanalyse die a-priori-Kategorisierung der Fragestellung bestätigen würden (Konfirmatorischer Ansatz). Das bedeutet, daß Fragen, die aus gleichen Ausprägungen der beiden Kategorien

resultieren, im Durchschnitt ähnlicher beurteilt werden als Fragen zu verschiedenen Ausprägungen der beiden Kategorien. Z.B. sollten Antworten zu Fragen der Zufriedenheit mit der Stadt als Wohnort und der Nachbarschaft im allgemeinen ähnlicher ausfallen als etwa Antworten zur derzeitigen Arbeit und Gesundheit im allgemeinen. Für eine Auswertung der Ergebnisse mit MDS heißt das, daß Fragen gleicher Kategorienausprägung räumlich zusammen liegen sollten, Fragen, die sich in nur einer der beiden Kategorien gleichen, weiter entfernt liegen sollten, Fragen zu ungleichen Ausprägungen weit voneinander entfernt wiederzufinden sein sollten.

Die bei der USA- und Israel-Studie erhobenen Daten bestätigten nicht nur die Vermutung, sondern ließen zusätzlich eine geometrische Aufteilung des Konfigurationsraums zu, der als "Radex"-Darstellung bezeichnet wird: Die Fragen mit Ausprägung "Zustand" der Facette A gruppieren sich in einem Kreis um den Schwerpunkt der Darstellung, die Fragen mit Ausprägung "Resource" der Facette A liegen außerhalb um diesen Kreis herum. Ferner läßt sich der Raum nach den Ausprägungen der Facette B in Segmente aufteilen.

INTERRELATIONSHIPS AMONG FIFTEEN VARIABLES* OF SATISFACTION WITH LIFE AREAS IN THE UNITED STATES



aus : (23)

In einer weiteren Studie (Levy and Guttman, (22)) wurde ein Design mit sechs zugrundegelegten Facetten benutzt, das eine MDS-Darstellung in vier Dimensionen erbrachte. Zur geometrischen Interpretation wurden Bezeichnungen wie "Zylindrex", "Multiplex", "Torus" etc. bemüht. Wieder erbrachte die Studie eine Bestätigung der vorher fixierten Facetten durch die Gesamtheit der Befragten.

In anderen Arbeiten wurde die Anwendung eines Verfahrens ("CMDA") empfohlen, das die MDS-Lösung von vornherein in eine mit der Ausgangshypothese kompatible Konfiguration zwingt, d.h. dem MDS-Verfahren Beschrän-

kungen so auferlegt, daß eine konfirmatorische Lösung erzielt wird. Das Streßmaß dient dann zur Überprüfung, inwieweit die Lösung auch mit den Ausgangsdaten übereinstimmt. Auf diese Weise lassen sich Abweichungen von der konfirmatorischen Lösung, die rein verfahrenstechnisch oder zufallsbedingt sind, vermeiden.

Die Relevanz des Facettenansatzes soll hier keineswegs bestritten werden. Ergebnisse der Art "je ähnlicher der Themenkomplex, desdo ähnlicher die Antworten im Durchschnitt der Befragungen" sind eine Bestätigung für den Forscher, daß die von ihm in die Untersuchung eingebrachten Kategorien auch vom Gros der Befragten für relevant befunden werden. Facetten-theorie ermöglicht damit eine Validitätsprüfung des Forschungsdesigns oder die Prüfung von Hypothesen über Facetten.

Die MDS-Verfahren liefern allerdings anschauliche Grafiken für diesen Kontext, und die Ästhetik schlauchförmiger Gebilde ist anscheinend gut geeignet, zu Spekulationen über Strukturen gesellschaftlichen Bewußtseins beizutragen. Guttman ging sogar einen Schritt weiter, indem er bei Anwendung des "powerful facet theoretic approach" "Gesetze" menschlichen Verhaltens zu entdecken vermochte (Levy (24)). Die erwähnten konfirmatorischen MDS-Ergebnisse werden als "zweite Gesetze" bezeichnet, die mit der empirischen Bestätigung der "regionalen Hypothesen" (das sind die Strukturen der MDS-Konfigurationen) legitimiert wird.

Die "ersten Gesetze" erstrecken sich über die Gebiete Intelligenz, Einstellung (attitude) und Bezug (involvement). Da sie alle ähnlich lauten, sei hier eines für alle zitiert: Zunächst die Definition von Einstellung (attitude):

Ein Aspekt gehört zum Universum der Einstellungsaspekte

dann und nur dann,	}	kognitiver
wenn er nach		affektiver
		instrumenteller Haltung

in bezug auf ein Objekt fragt, und die Antwortmöglichkeit von sehr positiv bis sehr negativ rangiert (11).

Das erste Gesetz der Einstellung:

"Wenn zwei beliebige Aspekte aus dem Universum der Einstellungsaspekte zu einem gegebenen Objekt ausgesucht werden, und wenn die beobachtete Population nicht künstlich ausgesucht wurde, dann wird die Regression zwischen den beiden Aspekten von positivem Anstieg oder mit Anstieg gleich null in dieser Population sein." (11).

Das heißt also vereinfacht, daß eine Durchschnittsbevölkerung alle Aspekte zu einem Gegenstand im Durchschnitt mit gleicher Tendenz oder unabhängig beurteilt. Wenn eine Sache gut ist, dann ist sie, über alle Meinungen gemittelt, in keiner ihrer Eigenschaften schlecht.

Guttman's Gesetze werden nicht bewiesen, sondern "bestätigt". Levy (24) führt sechs Studien an, die das Gesetz der "Einstellung" stützen. Zitiert sei hier die Tabelle aus der Untersuchung über das persönliche Wohlbefinden von Studenten in nicht-religiösen Schulen (24). Tatsächlich weisen fast alle ausgesuchten Aspekte eine positive "Korrelation" (hier ausgedrückt im Monotonie-Koeffizienten, vgl. Kapitel 3) miteinander auf, die restlichen weichen zumindest nicht sehr stark negativ von Null ab.

Shlomit Levy
Louis Guttman

Monotonicity coefficients among values for personal wellbeing
for students in secular schools

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Learning achievements	1	--	40	60	47	38	24	23	17	43	08	41	11	34	44	33	31	32
Achievements in sports	2	40	--	38	28	39	22	30	23	15	20	40	38	32	31	28	29	20
Attractive appearance	3	60	38	--	82	61	66	28	47	50	16	22	21	22	25	50	26	32
Fashionable clothes	4	47	28	82	--	82	59	21	43	33	05	30	11	12	31	48	21	22
Money for entertainment	5	38	39	61	82	--	56	25	49	20	03	33	15	13	36	46	19	14
Success with opposite sex	6	24	22	66	59	56	--	24	69	47	14	-02	24	13	-01	50	11	13
Social activity	7	23	30	28	21	25	24	--	51	40	31	40	27	46	36	41	70	50
Spend time with friends	8	17	23	47	43	49	69	51	--	64	20	04	34	19	04	64	34	35
Good family relations	9	43	15	50	33	20	47	40	64	--	36	12	34	48	08	54	44	56
Chance to develop talents	10	08	20	16	05	03	14	31	20	36	--	28	59	29	15	08	32	30
Biblical study	11	41	40	22	30	33	-02	40	04	12	28	--	37	48	91	23	54	47
Engage in hobbies	12	11	38	21	11	15	24	27	34	34	59	37	--	47	24	30	34	34
Contribute to State	13	34	32	22	12	13	13	46	19	48	29	48	47	--	39	30	67	54
Be observant	14	44	31	25	31	36	-01	36	04	08	15	91	24	39	--	23	53	45
Be one of the 'gang'	15	33	28	50	48	46	50	41	64	54	08	23	30	30	23	--	49	39
Participate in public assignments	16	31	29	26	21	19	11	70	34	44	32	53	34	67	53	49	--	67
Help others	17	32	20	32	22	14	13	50	35	56	30	47	34	54	45	39	67	--

aus: (22)

Bleibt zu fragen, was sich denn in solchen positiven Korrelationen ausdrückt. Sind das wirklich "Gesetze" der Einstellung, sind das nicht vielleicht Beispiele für einen - über alle gemittelten - Harmoniezustand in bezug auf das Objekt der Einstellung? Was geschieht mit dem "Gesetz", wenn sich die Einstellung zu einigen Aspekten des Objekts dramatisch verändert? Nach Guttman wäre das nicht möglich, d.h. es wird einfach die Möglichkeit der Existenz einer widersprüchlichen oder auch nur differenzierten Einstellung für eine Gesamtbevölkerung bestritten. Welche Rolle spielt die Erhe-

bungssituation, die Anhäufung von Fragen zum selben Komplex in einem Interview? Ganz zu schweigen von der Relevanz dieses Gesetzes, das sich ja auf eine Gesamtpopulation bezieht und über Teile oder Individuen überhaupt keine Aussage machen kann.

Ungeachtet dieser offenbleibenden Fragen werden hier "Gesetze" der Sozialwissenschaften formuliert, in Anlehnung an die Gesetze der Naturwissenschaften (14). Demnach ist alles das bereits "Gesetz" in den Sozialwissenschaften, dem nach einigen empirischen Untersuchungen nicht unbedingt widersprochen werden kann - die empiristische Vorstellung vom Begriff des Gesetzes. Daß die angeführten empirischen Studien ihre Ergebnisse im Forschungsdesign schon implizieren und unter den eingegangenen Forschungsbedingungen die Ergebnisse beinahe zwangsläufig erbringen müssen, schließt den Kreis: Die Grenze zur Tautologie ist hier überschritten.

Hinter einem solchen Forschungsansatz und -ergebnis steht im übrigen eine angemessene Erkenntnistheorie. Guttman selbst sprach auf einem Symposium (21) den Satz aus: "We don't need any scientific axioms, the data are our axioms.". Die Daten sind unsere Axiome - man möchte hinzufügen: Man muß nur empirisch geschickt genug agieren, und schon fallen von den erhobenen "Datenaxiomen" die Gesetze der Sozialwissenschaften ab.

Interessant in diesem Kontext ist auch Guttman's Definition des Begriffs "Theorie" (11): "Eine Theorie ... ist eine Hypothese über eine Korrespondenz zwischen einem Definitionensystem für ein Universum von Beobachtungen und einem Aspekt der empirischen Struktur dieser Beobachtungen, zusammen mit einer Be-

gründung für solch eine Hypothese.". Levy (24)
schließt an dieses Zitat an: "Das impliziert, daß der
Forscher sich um zwei verschiedene Dinge kümmern muß:

1. das Design seiner Beobachtungen und
2. die empirische Struktur jener Beobachtungen.

Der theoretische Sprung verlangt das Denken über beide
Dinge im Zusammenhang.". Weiter schreibt sie: "Das Fa-
cettendesign liefert das Definitionensystem der Beo-
bachtungen zusammen mit der Begründung für Hypothesen."

Die Definition läßt allerdings mit der "Begründung
für eine solche Hypothese" einem Rückzug womöglich
in rationalistische Argumentation Tür und Tor offen.
Interessant bleibt aber, daß "Theorie" selbst zur Hy-
pothese degradiert wird, die auf wie auch immer "be-
gründete" Art und Weise die Lücke zwischen einem (me-
taphysischen? universalen?) Definitionensystem und
der Empirie schließt.

Somit entstehen aus dieser Definition mehr Fragen,
als sie an Klarstellungen eröffnet. Der Beliebigkeit
des Definitionensystems sowie der Begründung der Hy-
pothesen sind keine Grenzen gesetzt. Das Definitio-
nensystem ist dem "Universum der Beobachtungen", was
immer das sein soll, unmittelbar aufgesetzt. Es bleibt
offen, was eine "Korrespondenz" zwischen Definitionen-
system und einem Aspekt der empirischen Struktur aus-
macht; was z.B. überprüft die empirische Forschung?

Levy vereinfacht die Fragestellung durch die Beschrän-
kung auf Forschungsdesign und empirische Struktur der
Beobachtungen, also auf die Empirie selbst. Ins Design
wird gleich noch Definitionensystem und Begründung
mit hineingenommen. Vielleicht kann man zur weiteren
Vereinfachung vorschlagen, auch Design und Beobachtun-
gen als identisch anzunehmen, wie oben erörtert, dann

hätte man neben der Einfachheit auch die Klarheit über das Zustandekommen der Ergebnisse.

Die ausführliche Diskussion dieses Ansatzes erschien hier deshalb gerechtfertigt, weil er symptomatisch für viele Protagonisten der Multidimensionalen Skalierung ist. Ohne daß sich alle bekannten MDS-Anwender und -Entwickler gerade Guttman's Vorgehensweise und Ideengut angeschlossen, zeigen doch viele einen ausgesprochenen Hang zum empiristischen Forschungsstil: "Laßt die Daten sprechen" ist die Maxime, "darstellen lassen sie sich anschaulich, eine Begründung / Interpretation ist dazu erstellbar und - was will man mehr - sogar konfirmatorisch - gesetzgeberisch kann gearbeitet werden".

Beschränkungen für die Anwendung der MDS -

Der Probabilistische Ansatz

Nicht alle sind so optimistisch über die Einsatzmöglichkeiten der MDS wie die oben zitierten Autoren. Tatsächlich sind die Meinungen über den Nutzen der MDS sehr unterschiedlich. Eine Reihe von Kritikpunkten geben u.a. Bauer (1), Bauer und Skarabis (2) sowie Zinnes und MacKay (48) an. Danach ist die MDS nicht geeignet, das Zustandekommen der komplexen Urteile von Probanden über die Objekte zu erklären. Ferner bleibt offen, wie der darzustellende "Wahrnehmungsraum" beschaffen ist, d.h. welche Metrik zugrundeliegt, inwiefern der Raum konsistent oder isotroph ist, und ob der Raum für einen Probanden von Urteil zu Urteil als konstant angenommen werden kann - wie es die MDS voraussetzt (1 , 2 , 48).

Da mit Hilfe der MDS in der Regel lediglich über die Antworten der Probanden Aussagen gemacht werden können, unterstützt das Verfahren eher behaviouristische Auffassungen in der Psychologie.

MDS-Modelle sind deterministischer Natur, d.h. schließen probabilistische Überlegungen aus (2 , 48). Ausnahmen sind die Arbeiten von Ramsay (29) sowie Zinnes und Griggs (47). In ihnen werden Verteilungsannahmen über den Wahrnehmungsraum der Probanden gemacht, die allerdings keineswegs plausibel sind. Insbesondere deshalb nicht, weil -

1. asymmetrischen Eigenschaften des Urteilsverhaltens nicht Rechnung getragen wird und -
2. die Abhängigkeiten der Distanzen untereinander wegen ihrer Komplexität nicht berücksichtigt werden

können.

Nicht zuletzt wegen des fehlenden probabilistischen Konzepts existiert kein Testverfahren für MDS-Darstellungen. MDS-Modelle können daher in der Regel nicht falsifiziert bzw. abgelehnt werden. Die Ablehnungsmöglichkeit eines Modells ist aber eine absolute Notwendigkeit, da es nur unter dieser Bedingung überhaupt eine Aussage über die Realität machen kann (48). Insbesondere kann nicht überprüft werden, ob die psychologische Vorstellung des "Wahrnehmungsraums" zutrifft, d.h. ob der psychischen Wahrnehmung von Objekten eine Darstellung im reellen Raum angemessen ist.

Für die soziologische Anwendung heißt das, daß nicht verfahrensimmanent geprüft werden kann, ob eine räumliche Darstellung den erhobenen Daten entspricht. Es existiert kein verfahrenstechnisches Hilfsmittel, das hinreichend genau signalisieren könnte, ob durch die MDS-Darstellung eine Interpretation der Daten entsteht, die den Beobachtungen an sich fremd ist. Mag das für Objekte, die aus dem gleichen Kontext stammen - wie etwa die Medikamente - ein zunächst redundanter Gesichtspunkt sein, so ist das durchaus nicht der Fall, wenn man eine Untersuchung zu unterschiedlichen Themen durchführt und diese dann gemeinsam skalieren will, etwa nach dem Beispiel komplexer Facettendesigns, die ja die räumliche Darstellung unterschiedlicher Kategorien erbringen sollen.

Das traditionelle Hilfsmittel für die Überprüfung einer Hypothese ist aber das statistische Testverfahren, ein gewissermaßen rationales Pendant zur Empirie. Rational insofern, als Testverfahren explizite "vernünftige" Annahmen über die Verteilung der Daten, also

über den Einfluß von Zufälligem, machen und von diesen Annahmen her eine Aussage darüber zulassen, ob die Abweichungen der Daten von der Hypothese so groß sind, daß die Daten nur mit geringer Wahrscheinlichkeit der genannten Hypothese entsprechen. Die Rationalität von Testverfahren liegt aber schon allein darin begründet, daß der Begriff des Zufalls überhaupt ins Verfahren eingeführt wird, da Zufall nicht empirisch definiert werden kann (vgl. Stegmüller (45)).

Gesteht man dem Zustandekommen von empirischen Beobachtungen einen Zufallseinfluß zu, so muß eine Möglichkeit im Analyseverfahren vorhanden sein, die die Zufälligkeit berücksichtigt. Zufall sei definiert als realitäts- oder erhebungsbedingte Abweichungsmöglichkeit der empirischen Daten von einer zugrundeliegenden Realitätsstruktur, die der Forschungsprozeß erkennen soll.

Die Verteilung muß dazu auf dem Raum der Beobachtungen definiert werden, muß also bei ordinalen Daten für deren Wertebereich definiert sein. Denn die Verteilungsannahmen über einen metrischen "Wahrnehmungsraum" oder Darstellungsraum implizieren schon eine der entscheidenden Hypothesen: die Skalierbarkeit; d.h. die Hypothese, ob die Daten überhaupt sinnvollerweise in einen metrischen Raum gehören, wird mit dieser Annahme grundsätzlich bejaht und steht prinzipiell nicht mehr zur Überprüfung frei.

In den folgenden Kapiteln wird daher nach der Diskussion der bisher entwickelten Koeffizienten und Tests für nichtmetrische MDS-Verfahren eine Möglichkeit gezeigt, unter Anwendung der Prinzipien der klassischen Teststatistik auf ordinalem Niveau Testverfahren zu verschiedenen Fragestellungen zu entwickeln. Explizit

angegeben werden Tests zu Einzelhypothesen sowie Tests über die Skalierbarkeit auf Grundlage verschiedener Verteilungsannahmen.

2. Kapitel

Verfahren der Nichtmetrischen MDS; Streßmaß

Ausgangspunkt der N-MDS sind Daten, die sich als Rangfolge von Unähnlichkeiten darstellen lassen und deren Informationsgehalt wesentlich ordinaler Art ist. Diese Unähnlichkeiten sollen mit δ_{ij} , $i, j = 1, \dots, n$ bezeichnet werden, wobei i und j für je eines der n Objekte stehen, zwischen denen die Unähnlichkeit gemessen wurde. Setzt man noch Indizes zur Bezeichnung des Rangplatzes, so läßt sich eine Rangfolge schreiben als:

$$(2.1) \quad F(\delta) := \delta_{i1,j1} \geq \delta_{i2,j2} \geq \dots \geq \delta_{im,jm}, \quad m = \frac{n(n-1)}{2}$$

Eine solche Rangfolge kann z.B. zustande kommen, wenn in einem Testverfahren eine Person die Unähnlichkeiten je zweier Paare von Objekten (A_i, A_j) und (A_k, A_l) vergleicht und angibt, welche der beiden zugehörigen Unähnlichkeiten δ_{ij} und δ_{kl} größer ist. Wird dieses Verfahren für alle der $\binom{n(n-1)}{2}$ Vergleiche von zwei

Unähnlichkeiten wiederholt, mindestens aber für $m-1$, so ergibt sich eine vollständige Ordnung auf $\{\delta_{ij}\}$.

Aufgabe der NMDS ist es nun, eine räumliche Darstellung der n Objekte so zu finden, daß die Rangordnung der Distanzen zwischen den Objekten der Rangfolge der Unähnlichkeiten möglichst entspricht. Bezeichnet man die Distanzen einer MDS-Darstellung der n Objekte mit d_{ij} ; $i, j = 1, \dots, n$ und ihre Rangfolge mit $F(d) := d_{i1,j1} \geq d_{i2,j2} \geq \dots \geq d_{im,jm}$, so soll die Isotoniebeziehung $\delta_{ij} \geq \delta_{kl} \leadsto d_{ij} \geq d_{kl}$ für möglichst wenige $(\delta_{ij}, \delta_{kl})$ verletzt werden - wobei die Dimensionalität des Raumes, in dem die Objekte

skaliert werden, möglichst gering ist.

Eine nach diesem Prinzip konstruierte, "optimale" MDS-Darstellung bildet also die gemessenen Unähnlichkeiten als räumliche Distanzen ab, wobei die Ordnungsrelation im "Bildraum" möglichst erhalten bleibt. Die MDS-Konfiguration dient der übersichtlicheren Darstellung des ordinalen Datenmaterials und soll eine Hilfe bei der Interpretation komplexerer Beziehungsstrukturen sein, die in der rein ordinal-numerischen Darstellung kaum zutage treten.

Verfahren

Kruskal (19) beschreibt ein Verfahren zur MDS, das sich lediglich auf die ordinalen Informationen des Datenmaterials stützt (die Monotoniebedingung erfüllt) und das Kriterium für den Vergleich zwischen Daten und MD-Skala ("Streß") als integralen Bestandteil enthält. Das Verfahren geht so vor, daß eine im grunde beliebige Ausgangskonfiguration von n Punkten schrittweise so verschoben wird, daß die Rangfolge der Distanzen die Isotonieforderung immer besser erfüllt.

Die Distanzen d_{ij} einer Konfiguration werden dabei in Richtung auf Vergleichsgrößen $\hat{d}_{ij}$ verschoben, die nach jedem Schritt erneut berechnet werden müssen. Die $\hat{d}_{ij}$ werden berechnet, indem zunächst eine isotone Regression der δ_{ij} auf die d_{ij} hergestellt wird und dann die d_{ij} auf die Regressionskurve projiziert werden. Die Projektionen heißen $\hat{d}_{ij}$.

stand der $\hat{d}_{ij}$ von den zugehörigen d_{ij} möglichst klein zu halten, verlangt man von den $\hat{d}_{ij}$, daß sie die Summe der quadratischen Abweichungen von den d_{ij} minimieren, d.h.

$$(2.2) \sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2 = \min_{x_{ij} \in \mathbb{R}} \sum_{ij} (d_{ij} - x_{ij})^2$$

Daher gilt hier:

$$\hat{d}_{i1j1} = \hat{d}_{i2j2} = \hat{d}_{i3j3} = \frac{1}{3} \sum_{k=1}^3 d_{ijkj_k}$$

Die $\hat{d}_{ij}$ entsprechen demnach nicht realen Distanzen in einer Konfiguration - d.h. in der vorliegenden Konfiguration können die d_{ij} in der Regel nicht in $\hat{d}_{ij}$ verschoben werden, da für die $\hat{d}_{ij}$ die Metrikaxiome in der Regel nicht erfüllt sind. Die $\hat{d}_{ij}$ dienen vielmehr als Richtgrößen für die Veränderung der d_{ij} . Sie müssen nach jedem Iterationsschritt für die veränderte Konfiguration erneut berechnet werden.

Das Verfahren der isotonen Regression und Verschiebung wird so lange wiederholt, bis eine optimale Anpassung an die Isotonieforderung erreicht ist.

Streß

(2.2) wird auch als "raw stress" bezeichnet. Es ist ein Maß für die "Entfernung" einer Konfiguration von der Isotonieanforderung.

Um den Einfluß großer Abweichungen nicht überzubewerten, wird noch eine Normierung vorgenommen zu

$$(2.3) \text{ stress}_1 = S(x_1, \dots, x_{np}) = \min_{\substack{\hat{d}_{ij}, \text{ die isoton} \\ \text{zu } \delta_{ij} \\ \text{sind}}} \sqrt{\frac{\sum_{ij} (d_{ij} - \hat{d}_{ij})^2}{d_{ij}^2}}$$

oder:

$$(2.4) \text{ stress}_2 = S_2(x_1, \dots, x_{np}) = \min_{\substack{\hat{d}_{ij} \text{ isoton} \\ \text{zu } \delta_{ij}}} \sqrt{\frac{\sum_{ij} (d_{ij} - \hat{d}_{ij})^2}{\sum_{ij} (d_{ij} - d_{..})^2}}$$

wobei $\begin{pmatrix} x_1 \\ \vdots \\ x_{np} \end{pmatrix}$ der Vektor der Koordinaten von n Punkten.

im R^P ist.

Die beiden Streßkoeffizienten geben für den gleichen raw stress unterschiedliche Ergebnisse. Streß₁ ist historisch (Kruskal) von größerer Bedeutung; Streß₂ hat Vorteile, wenn die Daten viele Unähnlichkeiten mit gleichem Rang (ties) enthalten.

Eine MDS-Konfiguration wird "optimal" genannt, wenn sie den Streß minimiert - der Streß ist das wesentliche Entscheidungskriterium für die Optimalität einer Konfiguration. Es gibt eine vielzitierte Bewertung der Streßergebnisse durch Kruskal (19), nach der ein Streß von unter 0.01 ausgezeichnet,

0.01 - 0.05 gut

0.05 - 0.10 befriedigend

0.10 - 0.15 mäßig

und über 0.15 schlecht sei.

Shepard (37) bezeichnete dies als fragwürdig. Sie

kann zur groben Beurteilung dienen, nicht jedoch Basis für ernsthaftes Hypothesentesten sein.

Gradientenverfahren

Das Streßmaß dient nicht nur als Maß des Abstands einer Konfiguration von der Monotoniebedingung, sondern ist ein integraler Bestandteil des Verfahrens insofern, als es die Richtung bei der schrittweisen Verschiebung der Konfiguration hin zu einer optimalen Lösung angeben kann. Unter einer optimalen Lösung soll dabei diejenige verstanden werden, für die gilt, daß sie die Streßfunktion $S(x_1 \dots x_{np})$ minimiert. Aufgabe der Lösungsversuche ist es also, das oder ein globales Minimum der Funktion $S(x_1 \dots x_{np})$ zu finden. Da S eine Funktion von $n \times p$ Variablen ist, ist die Lösung nur mittels eines iterativen Verfahrens zu finden. Kruskal schlägt das Verfahren des "steilsten Abstiegs", auch "Gradientenverfahren" genannt, vor.

Dazu wird zunächst eine (im allgemeinen beliebige) Anfangskonfiguration von n -Punkten im p -dimensionalen Raum gebildet und deren Streßmaß berechnet. Die nächst bessere Konfiguration ist dann in Richtung des negativen Gradienten der Funktion S an der Stelle der Anfangskonfiguration zu suchen. Die Konfiguration wird also ein Stück weit in diese Richtung verändert. Danach wiederholt sich das Verfahren, bis ein Minimum gefunden wird (vgl. Kruskal(19)).

Wichtig für das Vermeiden von nur lokalen Minima und ein Herausschießen über ein Minimum ist die Schrittweite, die die nächste Konfiguration von der vorigen entfernt ist. Sie wird berechnet nach einer Formel, die die vorangegangene Schrittweite sowie die Größe

des Gradienten (Summe der quadrierten Komponenten), den vorigen Gradienten und den Streß zum gegenwärtigen Zeitpunkt, zum vorigen Zeitpunkt und zum Zeitpunkt vor fünf Schritten berücksichtigt.

Guttman's Verfahren

Guttman entwickelte 1968 ein weiteres Verfahren, das sich von Kruskals im wesentlichen in zwei Punkten unterscheidet: Statt des Streßmaßes benutzt Guttman einen Entfernungs-Koeffizienten (Coefficient of alienation), und statt der Vergleichsgrößen $\hat{d}$, die mittels der monotonen Regression errechnet wurden, bevorzugt Guttman Vergleichsgrößen d^* , die "Rangbilder" (rank images).

Rangbilder

Schreibt man die Distanzen einer Konfiguration in aufsteigender Reihenfolge, so findet man das Rangbild d_{kl} einer Unähnlichkeit $\delta_{i_m j_m}$, indem man in der Rangfolge der Distanzen zur m -ten Position geht, d.h.

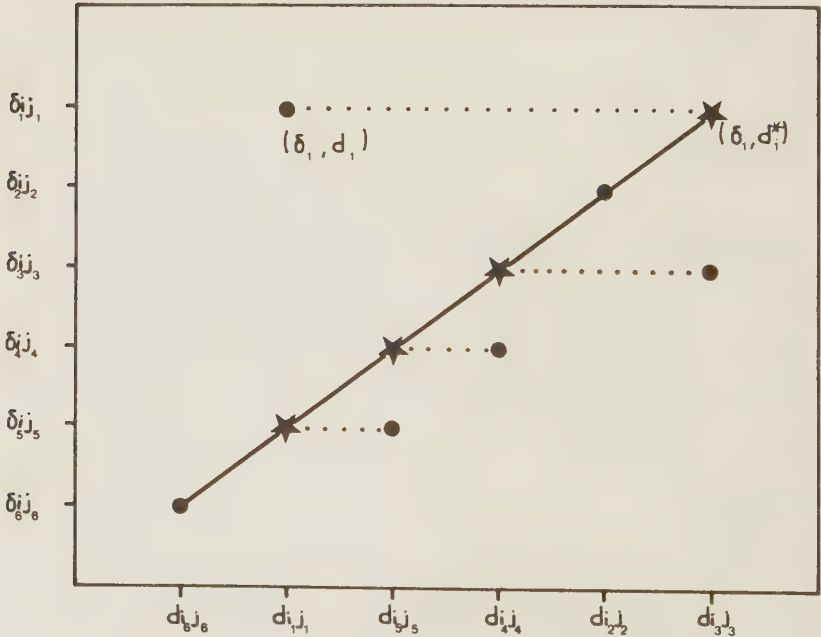
$$\text{rank image } (\delta_{i_m j_m}) = d_{ij}^* = d_{k_m l_m}.$$

Beispiel:

Sei $F(d) = \langle d_{12}, d_{23}, d_{13} \rangle = \langle 20, 30, 40 \rangle$ und $F(\delta) = \langle \delta_{12}, \delta_{13}, \delta_{23} \rangle$. Hier steht δ_{23} an dritter Stelle der Rangfolge der Unähnlichkeiten. Das Rangbild von δ_{23} ist daher gleich $d_{13} = 40$.

Wenn also Kruskal als Vergleichsgrößen die Mittelwerte aller Distanzen nimmt, die benachbart sind und das Monotoniekriterium nicht erfüllen, so werden bei Guttman keine Durchschnittswerte bestimmt, sondern die die Monotonieanforderung verletzenden d_{ij} so lange permutiert, bis die resultierenden d_{ij}^* die Bedingung erfüllen. Die d_{ij} einer Konfiguration werden so mit Werten aus der Folge d_{ij} selbst verglichen. Die d_{ij}^* minimieren daher im allgemeinen nicht den raw stress.

Graphik 2.2 : Beispiel einer rank-image Transformation



(Rangfolgen wie in Graphik 2.1)

Entfernungskoeffizient

Anstelle des Streßmaßes tritt der Koeffizient

$$K = \sqrt{1 - \frac{(\sum_{ij} d_{ij} d_{ij}^*)^2}{\sum_{ij} d_{ij}^2 \sum_{ij} d_{ij}^{*2}}} \quad (2.5)$$

K verhält sich monoton zu S und es gilt $K = S$, falls man statt d^* in (2.5) $\hat{d}$ einsetzt, da dann

$$K = \sqrt{\frac{\sum d_{ij}^2 \sum \hat{d}_{ij}^2 - (\sum (d_{ij} \hat{d}_{ij}))^2}{\sum d_{ij}^2 \sum \hat{d}_{ij}^2}}$$

$$= \sqrt{\frac{\sum \hat{d}_{ij}^2 \sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2 \sum \hat{d}_{ij}^2}} = \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}}$$

Ferner gilt $K = \sqrt{1 - (1 - (1/2 S^2))^2}$, falls d^* in S eingesetzt wird (30).

Die beiden Vorgehensformen von Guttman und Kruskal unterscheiden sich also nicht wesentlich, können vielmehr als Varianten ein und derselben Grundidee aufgefaßt werden. Da Guttman's Verfahren schneller zu einer Lösung führt und eher lokale Minima vermeidet, Kruskals Verfahren aber eine feiner angepaßte Lösung erzeugt, werden im Programm MINISSA beide Verfahren nacheinander durchgeführt. (vgl. (30))

Gebundene Werte, drei Arten der Monotonie

"Ties" oder gebundene Werte sind Unähnlichkeiten mit gleichem Rangplatz. Enthält das Datenmaterial gebundene Werte, so stellt sich die Frage, wie sich die entsprechenden Distanzen zueinander verhalten sollen. Verlangt man, daß

$$a) \delta_{ij} \leq \delta_{kl} \longrightarrow \hat{d}_{ij} < \hat{d}_{kl} \text{ und}$$

$$\delta_{ij} = \delta_{kl} \longrightarrow \hat{d}_{ij} = \hat{d}_{kl}$$

(gebundene Werte bleiben gebunden als Distanzen; strenge Monotoniebedingung, siehe Guttman (13))

oder lediglich daß

$$b) \delta_{ij} < \delta_{kl} \longrightarrow d_{ij} < d_{kl}$$

(gebundene Werte können gelöst werden; abgeschwächte Monotoniebeziehung)

oder aber daß

$$c) \delta_{ij} < \delta_{kl} \longrightarrow \hat{d}_{ij} \leq \hat{d}_{kl}$$

(nichtgebundene Werte können gebunden werden; schwache Monotoniebeziehung),

so ergibt sich bezüglich der Anzahl der unterschiedlichen Rangplätze (Rangzahl) der Rangfolgen $F(\delta)$ und $F(\hat{d})$, daß sie bei

a) gleich sind, bei

b) Rangzahl $(F(\delta)) \leq$ Rangzahl $(F(\hat{d}))$ und

c) Rangzahl $(F(\delta)) \geq \text{Rangzahl}(F(\hat{d}))$ ist.

Die strenge Monotoniebedingung erfordert unter Umständen die Darstellung in einem Raum höherer Dimensionalität. Eine Darstellung von n Punkten bei schwacher Monotonie ist z.B. in einem Raum der Dimension $n - 2$ immer mit Streßwert 0.0 möglich. Bei starker Monotoniebedingung benötigt man bis zu $n - 1$ Dimensionen. Voraussetzung ist natürlich, daß die Unähnlichkeiten ordinaler Natur sind.

Metrik

Als Metrik des Konfigurationsraumes dient zumeist die Euklidische oder eine andere der L_p -Metriken, auch Minkowski-Metriken genannt. Shepard (37) gibt einige weitere Möglichkeiten an, Shepard und Cunningham (7) entwickelten ein Verfahren der Transformation der Unähnlichkeiten in Größen, die lediglich die Metrikaxiome erfüllen und somit im allgemeinen nicht als räumliche Distanzen darstellbar sind (Nonspacial oder nondimensional NMDS). Minkowski Metriken haben die Eigenschaften der Linearität, d.h.

$$d(x, y) = d(x + z, y + z) \text{ und}$$

$$d(kx, ky) = k \cdot (d(x, y)).$$

Die Euklidische Metrik hat neben der Eigenschaft, daß in ihr die Distanzen visuell leicht einzuschätzen sind, auch den Vorzug gegenüber den anderen L_p -Metriken, daß in ihr Konfigurationen rotationsinvariant sind.

Räume mit unterschiedlicher Metrik erzielen MDS-Konfigurationen mit unterschiedlichem Streß. Die Wahl der

Metrik scheint dennoch nur aus inhaltlichen Erwägungen begründbar zu sein (z.B. City-Block-Metrik bei Darstellung von Variablen aus Kontingenztafeln) und nicht aus Gründen der Streßminimierung.

Dimensionalität

Viele Anwender versuchen, eine optimale oder auch die wahre zugrundeliegende Dimension ihrer Daten zu finden. Die MDS-Computerprogramme haben jedoch keine Möglichkeit zu entscheiden, welches die beste Dimensionalität des Konfigurationsraumes ist. Allerdings ist die Größe des Streßes einer Konfiguration abhängig von der Anzahl der Dimensionen des Raumes, in den die Objekte abgebildet werden. In der Tat ist es möglich, einen Raum zu finden, in dem die Unähnlichkeitsurteile isoton in Distanzen zu überführen sind. Dieser Raum hat maximal $n - 1$ Dimensionen.

Mit wachsender Dimensionalität des Raumes muß der Streß sinken. Kruskal und Guttman schlagen vor, die Werte ihrer Streßmaße für Darstellungen in mehreren Dimensionen zu beobachten. Eine Dimension gilt dann als optimal, wenn durch Hinzufügen einer weiteren Dimension die Anpassung (bzw. der Streß) sich nicht wesentlich verbessert. Kruskal spricht vom "Ellbogen" der Funktion Streß (Dimension des Darstellungsraumes), der die optimale Dimension kennzeichnet. Neben dieser groben Regel wird jedoch als Entscheidungskriterium letztlich auf eine substantielle Theorie verwiesen, die die Dimensionalität plausibel interpretieren muß.

Ungeklärt ist ferner die Frage, wie viele Punkte notwendig sind, um die Darstellung in einer bestimmten Anzahl von Dimensionen überhaupt zu ermöglichen. De-

generierte Messungen können entstehen, wenn die Anzahl der Punkte zu klein ist im Verhältnis zur Anzahl der Dimensionen (z.B. Schubert,⁽³⁴⁾). Spence hat folgende Mindestpunktzahl in den Dimensionen 1 - 3 vorgeschlagen:

wenigstens 6 Punkte für eine Dimension
 wenigstens 10 - 11 Punkte für zwei Dimensionen
 wenigstens 17 - 18 Punkte für drei Dimensionen.

Falls wiederholte Messungen vorliegen oder bei metrischer MDS können auch jeweils weniger Punkte ausreichend sein. Seine Faustregel entstand unter Anwendung von Monte-Carlo-Studien.

1974 versuchten Spence und Graef mit Hilfe von Monte-Carlo-Studien Vergleichswerte zu finden, die eine bessere Einschätzung der wahren Dimensionalität einer nichtmetrischen MDS-Lösung ermöglichen. Sie nahmen dazu zufallsgenerierte, 1 - 5 dimensionale Punktekongfigurationen mit 12 bis 36 Punkten. Die Punkte der verschiedenen Konfigurationen waren relativ gleichverteilt im Raum (keine Cluster). Zu euklidischen Distanzen der Konfigurationen wurden dann nacheinander Fehlergrößen von fünf verschiedenen Graden addiert. Für jede Ausprägungskombination der Kategorien "Anzahl", "Dimensionalität" und "Fehlergrad" wurden mehrere nach Zufallsprinzip ausgewählte Konfigurationen erzeugt. Der Erwartungswert des Stresses dieser Konfigurationen ging dann in die Tabelle ein. Der Benutzer kann sich nun dieser Daten bedienen, indem er die Streßwerte der eigenen Daten in 1 - 5 Dimensionen berechnet und diesen Vektor mit den Streßvektoren von Spence - Graef vergleicht, denen die gleiche Punktzahl zugrundeliegt. Der Vergleich erleichtert die Entscheidung für eine bestimmte Dimensionalität und auch für

den Grad des Fehlers, mit dem die Daten behaftet sind: Spence und Graef empfehlen, den zufallsgenerierten Streßvektor zu betrachten, der dem Streßvektor der eigenen Daten am nächsten kommt - die zugrundeliegende Dimension und Fehlergröße des Zufallsverfahrens sei dann auch für die realen Daten anzunehmen. (vgl. (41))

Problematisch bei diesem Verfahren bleibt die Gebundenheit der Spence-Graef-Tabellen an die Voraussetzungen der Monte-Carlo-Studie: längs der "wahren" Dimensionen normalverteilte Fehlergrößen, Punktekongfigurationen ohne Cluster oder Ausreißer, euklidische Distanz. Anwendungen zeigen dann auch oft große Differenzen zwischen den Streßvektoren der Daten und den nächstliegenden zufallsgenerierten (vgl. Coxon und Jones, 1980 (6)), für die "die Methode nicht so aufgefaßt werden sollte, als gäbe sie Zutritt zu versteckten Wahrheiten". Im Zweifel ist es immer besser, sich für eine Konfiguration höherer Dimensionalität zu entscheiden.

Bewertungsmöglichkeiten

Als einziges Maß zur Bewertung von NMDS-Konfigurationen muß eines der Streßmaße dienen, das zudem oft die Aufgabe hat, die optimale Dimensionalität und Metrik des Bildraumes festzulegen. Das Streßmaß ist aber von seiner Konzeption und Anwendung her dazu nicht in der Lage. NMDS-Ergebnisse sind deshalb statistisch unzureichend abgesichert, was das NMDS-Verfahren oft inakzeptabel macht. Es seien hier einige der Probleme genannt:

- Das Streßmaß ist definiert auf dem Bildraum R^D der Konfigurationen und dort (fast sicher) stetig und

differenzierbar. Die Ausgangsdaten sind jedoch bei der nichtmetrischen MDS ordinal. Die Anpassung der Konfiguration an die Ausgangsdaten wird also auf viel feinerem Meßniveau gemessen, als die Ausgangsdaten. Dieses zunächst theoretisch anmutende Problem enthält einige wichtige Implikationen:

- a) Der Streß mißt nicht die Abweichung der Rangfolge der Distanzen einer Konfiguration von den Unähnlichkeiten wie es etwa ein ordinales Maß (Kendall's Tau, Spearman's rho) leisten könnte. Er mißt die Abweichung von einer (nächstliegenden) Folge von metrischen Vergleichsgrößen, die dem Monotoniekriterium entsprechen. Nächstliegend bedeutet dabei, daß der Streß von den Vergleichsgrößen minimiert werden muß. Was auf den ersten Blick plausibel und wünschenswert aussieht, kann aber gerade bei Konfigurationen mit wenigen Punkten erhebliche Konsequenzen mit sich bringen: Dann nämlich, wenn die relative, d.h. metrische Entfernung der Vergleichsgrößen klein ist, aber die Verletzung der Monotonieanforderung erheblich.
 - b) Es können sich bezüglich der Konfiguration Lösungen ergeben, die suboptimal bezüglich des Monotoniekriteriums sind, d.h. Lösungen, die aufgrund ihrer metrischen Nähe zu monotonen Vergleichsgrößen in rein ordinalem Maßstab weiter von den Ausgangsdaten entfernt sind, als Lösungen mit eventuell geringerem Streß.
- Die Lösung ist mit einer einzigen Konfiguration gegeben. Vorstellbar sind aber in vielen Fällen mehrere Konfigurationen, die das Monotoniekriterium nicht stärker verletzen und daher ebenso in Betracht kommen (5).

- Der Streßwert selbst liegt auf einer Skala, für die keine Akzeptanzbereiche, kritischen Bereiche etc. angebar sind. Die Einschätzung des Streßwerts geschieht "über den Daumen".
- Die MDS-Konfigurationen können nicht als statisch in dem Sinn betrachtet werden, daß die Orte der einzelnen Punkte durch die Ausgangsdaten festgelegt werden könnten. Vielmehr muß man oft davon ausgehen, daß die Punkte wechselseitig verschiebbar sind, ohne daß dadurch der Streß verringert, zumindest aber ohne daß dadurch das Monotoniekriterium stärker verletzt würde. Wegen der Interdependenz aller Distanzen ist aber eine Art Verteilungsannahme eines Punktes in R^D mit wachsender Punktezahl unmöglich. Es konnte bisher nicht einmal festgestellt werden, bis zu welchem Grade eine Konfiguration bei gegebener Punktezahl und Dimension festliegt.
- Von einigen Autoren wird eine MDS-Konfiguration als modellhafte Darstellung der Ausgangsdaten betrachtet (50). Für dieses Modell kann aber nicht die Null-Hypothese der Akzeptierung geprüft werden. Noch viel weniger ist es möglich, die Fülle der Gegenhypothesen, d.h. sämtliche anderen möglichen Darstellungsweisen, zu überprüfen. Dabei verliert gerade die Null-Hypothese für Modellüberprüfung im Vergleich zu traditionellen Testverfahren an Bedeutung, weil es hier nicht so sehr darauf ankommt, ein "richtiges" Modell nicht zu verwerfen, sondern gerade zu legitimieren, daß ein Modell vor allen anderen bevorzugt wird. Hierzu fehlen die Hilfsmittel, die das Verfahren bieten könnte. Entscheidungen müssen "substanzwissenschaftlich" begründet sein.
- So sehr substanzwissenschaftliche Überlegungen bei

statistischen Verfahren vonnöten sind - und auch in der "traditionellen" Statistik immer benötigt wurden - so sehr sind numerische Verfahren notwendig, um probabilistische Hilfestellung bei der Einschätzung der Ergebnisse zu leisten. Anderenfalls würde ein wichtiges Hilfsmittel voreilig über Bord geworfen, das ein Schutz gegen willkürliches Interpretieren bzw. plausibel machen unerwarteter Ergebnisse sein kann. Damit wird nicht der übergeordnete Wert substanzwissenschaftlicher Theorie bestritten, der auch bei "Nichtsignifikanz" eine Hypothese beibehalten läßt oder sich willkürlichen Interpretationen entgegensetzt. Andererseits ist die Frage, was denn durch bloße Visualisierung von Teilen der Information, die das Datenmaterial trägt, eigentlich für den Forschungsprozeß gewonnen ist. Dies um so mehr, wenn sich die Güte der Abbildung und der Anteil der "harten" Information nur unzureichend abschätzen läßt. Eine verschwommene Einsicht in empirische Daten ist aber schwer in klare und konsistente Theorieentwicklung einzubeziehen. Ferner kann sie auch für konfirmatorische Arbeiten nur von zweifelhaftem Wert sein.

3. Kapitel

Entscheidung zwischen MDS-Konfigurationen

Skarabis (39), sowie Lingoes und Borg (3) beschreiben MDS-Verfahren, mit deren Hilfe man Konfigurationen mit bestimmten Beschränkungen erzeugen kann. Die Verfahren bewirken z.B. die Fixierung einzelner Punkte im Konfigurationsraum, Fixierung von Distanzen oder Verhältnissen von Distanzen zueinander sowie von Winkeln. Diese Verfahren ermöglichen, Vorwissen über Konfigurationen in die Analyse mit einzubeziehen und dadurch Konfigurationen leichter interpretierbar zu machen. Außerdem kann überprüft werden, ob apriori Vorstellungen über die MDS-Lösung mit den Daten kompatibel sind, indem geprüft wird, ob auch die MDS-Lösungen mit den verlangten Beschränkungen den Daten hinreichend angepaßt sind.

Neben dem Streßmaß werden daher hier noch andere Indikatoren benötigt, die anzeigen sollen, ob die optimale unbeschränkte MDS-Konfiguration entscheidend von einer beschränkten Konfiguration abweicht, oder ob die Beschränkung keine wesentliche Änderung bewirkt. Zur Bewältigung dieses Problems wurden verschiedene Testverfahren (Hotelling's t , Kendall's Tau, Spearman's ρ) und Anpassungsmaße ("measures of fit") verwendet.

3.1 Hotelling's t

Man versucht hier eine lineare Regression der Unähnlichkeiten auf unterschiedliche MDS-Konfigurationen. Mit y werde die abhängige Variable bezeichnet, $x_1, \dots, x_q$ seien die Prädiktoren. Meßwerte für y sind hier also die Unähnlichkeiten, während Meßwerte für

x_i , $i = 1, \dots, q$ die den Unähnlichkeiten zugeordneten Distanzen der i -ten MDS-Konfiguration sind.

Lingoes und Borg normieren die x_i zu z_i , die den Mittelwert 0 und die Varianz 1 haben. Aus den z_i werden ferner orthonormierte Werte u_i gebildet, mit

$$u_i = \sum_j r^{ij} z_{ij}, \quad j = 1, \dots, w \quad (\text{Index für die Distanzen in einer Konfiguration}),$$

(r^{ij}) ist die Inverse der Korrelationsmatrix der Prädiktoren, $(r(z_i, z_j))^{-1}$.

Es wird dann die Regression von y auf die u_i berechnet. Die durch die u_i erklärte Varianz ist gleich der Varianz, die durch die x_i erklärt würde.

Man erhält:

	Anzahl der <u>Freiheitsgrade</u>
Regression auf u_1	= 1
Regression auf $u_2 \dots u_q$	= $q - 1$
Residuen	= $n - q - 1$
	$n - 1$

Wegen der Orthogonalität der u_i erhält man die von den $u_2, \dots, u_q$ erklärte Varianz als Differenz zwischen der von $x_1, \dots, x_q$ und der von u_1 erklärten. Wäre der F-Test hier anwendbar und würde, auf $u_2, \dots, u_q$ und die Residuen angewandt, nicht wesentlich von 1 abweichen, so könnte man schließen, daß keine signifikanten Differenzen zwischen den q Prädiktoren bestehen.

Hat man nur zwei MDS-Konfigurationen, x_1 und x_2 , so ist die Nullhypothese $\beta_1 = \beta_2$, d.h. Prädiktoren x_1

und x_2 sind gleich wirksam zur Erklärung von y . Normiert man wieder x_i zu z_i , so erhält man die einfachen Regressionskoeffizienten von y auf z_1 und z_2 durch

$$b_k = \frac{\sum_i y_i z_{ki}}{\sum_i z_{ki}^2}, \quad k = 1, 2 \quad i = 1, \dots, n.$$

Angenommen, z_1 und z_2 korrelieren positiv (anderenfalls wird eine der Variablen mit -1 multipliziert), so gilt

$$b_1 - b_2 = \frac{\sum_i y_i (z_{1i} - z_{2i})}{\sum_i (z_{1i} - z_{2i})^2}.$$

Hält man die z_{ki} 's bei wiederholten Stichproben fest, so ist die Varianz von $b_1 - b_2$ zu schätzen durch

$$\frac{\sum_i (z_{1i} - z_{2i})^2}{\sum_i (z_{1i} - z_{2i})^2} s^2, \quad s^2 = \text{Schätzung der Residualvarianz von } y \text{ nach der Regression auf } z_1 \text{ und } z_2.$$

Die Differenz $b_1 - b_2$ kann dann mittels t-Test mit $n - 3$ Freiheitsgraden getestet werden. (vgl. (27))

Die Voraussetzungen für das Testverfahren sind jedoch derart, daß es nicht auf die Daten der MDS-Analyse wie oben beschrieben angewendet werden dürfte. Nach Hotelling (16) sind zumindest folgende Annahmen einzuhalten:

- Die abhängige Variable y ist normalverteilt (univariat) für alle Mengen von (Meß-)werten von x_1 und x_2 , und zwar
- ... unabhängig für verschiedene Mengen $\{x_1, x_2\}$.
- y hat gleiche Varianz für alle $\{x_1, x_2\}$.
- $E(y)$ ist für festes $\{x_1, x_2\}$ eine lineare Funktion dieser Prädiktoren.

Unter Bedingungen a - d ist der Signifikanztest t anwendbar für die "Subpopulation aller möglichen Stichproben, für die die Prädiktoren x_1 und x_2 dieselben Werte haben wie in dem speziellen, gemessenen Beispiel" (16).

Nun ist von den vier Bedingungen keine bei MDS-Daten erfüllt. Lingoes und Borg halten den t-Test trotzdem für anwendbar, weil sie in umfangreichen Monte-Carlo-Studien eine erhebliche Robustheit des Verfahrens gegenüber Verletzungen der Annahmen feststellen konnten.

Zunächst bildeten sie die Werte der Unähnlichkeiten und der Distanzen auf ihre Rangplätze ab, d.h. in die

Menge $\{x = 1, \dots, \frac{n(n-1)}{2}\}$. Diese Transformation ist geeignet,

die Beziehung zwischen den Werten δ_{ij} und den d_{ij} zu "linearisieren" (27), d.h. stellt unter allen möglichen Transformationen für ordinale Daten eine derjenigen dar, die die Linearitätseigenschaften zwischen δ_{ij} und d_{ij} am wenigsten verletzen.

Die Monte-Carlo-Studie bestand aus der Generierung von Zufallswerten für y und x_1, x_2 , wobei verschiedene Verteilungsannahmen für y, x_1 und x_2 , sowie unterschiedliche Grade der Korrelation $r_{x_1 x_2}$ angenommen

wurden. Für jede Kombination der vorher festgesetzten Werte wurden 1.000 Zufallsstichproben mit je 45 Beobachtungen (10-Punkte Konfigurationen) gezogen und daraus die empirische Verteilungsfunktion für t ermittelt. Da die $r_{x_1 x_2}$ nichts zur Varianz innerhalb

der Parameter der empirischen Verteilungsfunktion beitrugen, konnte über sie summiert werden, so daß 5.000 Zufallsstichproben für jede der Kombinationen der fest-

gesetzten Verteilungen von x und y zur Verfügung standen. Die empirische Verteilungsfunktion von t unterscheidet sich dann kaum von der theoretischen.

Obwohl hier recht eindrucksvoll die Robustheit von Hotelling's t nachgewiesen wird, bleiben doch Zweifel bezüglich seiner Anwendbarkeit in der NMDS:

- Es handelt sich hier um eine große Zahl von Zufallsstichproben, so daß man nicht ausschließen kann, daß die Robustheit nur für große Stichproben gilt - NMDS hat es aber meist nur mit einer Stichprobe zu tun.
- y , d.h. $\langle \delta_{ij} \rangle$ unterliegt keiner speziellen Verteilungsfunktion, die Abhängigkeit der Komponenten von y ist vermittelt über die Abhängigkeit der Objekte, zwischen denen die Unähnlichkeiten gemessen werden. Inwiefern auch für sehr untypische - vom Standpunkt der "Verteilungsannahme" aus gesehen - Datenstrukturen die Robustheit noch besteht, ist nicht klar.
- Die Linearitätsbeziehung zwischen zwei Folgen von Rangplätzen dürfte erheblich stärker verletzt sein als vergleichbare kontinuierliche Wertemengen, der t -Test wäre daher immer geneigt, die gefundenen Differenzen überzubewerten.
- Schwierigkeiten mit Multikollinearität ergeben sich, wenn x_1 und x_2 hoch korrelieren, wie es ja in den wichtigen Fällen der konfirmatorischen Analyse geradezu gewünscht wird.

3.2 Intervallskalierte "Measures of Fit"

Schönemann und Carroll (33) sowie Lingoes und Schö-

nemann (25) entwickelten einige intervallskalierte Anpassungsmaße für Matritzen, mit denen der Anpassungsgrad von zwei unterschiedlichen MDS-Distanzmatritzen gemessen werden kann. Die Maße beziehen sich auf die Residuen E einer linearen Regression

$$(3.1) \quad B = c A T + I \gamma' + E,$$

wobei A und B die beiden zu untersuchenden $(p + q)$ -Matritzen sind. T ist eine $(p + q)$ -Rotationsmatrix, I ein $(p, 1)$ -Einheitsvektor und γ ein $(1, q)$ -Translationsvektor, c eine Konstante für Kontraktion/Dilatation. In diesem Modell wird also A durch Rotation, Translation und Stauchung oder Streckung so verändert, daß B optimal angepaßt wird. Die (p, q) -Residualmatrix E dient dann als Grundlage zur Messung der für die MDS-Problematik wesentlichen Abweichungen zweier Distanzmatritzen.

Ein grobes Maß der Anpassung ist die Frobeniusnorm der Matrix E, wobei E aus dem obigen Regressionsproblem bestimmt wird als $E = Q (B - c A T)$,

$$Q = I_{p,p} - JJ'/p, \text{ also}$$

$$(3.2) \quad e = |E| = \text{tr } E'E = \text{tr } B'QB - (\text{tr } T'A'QB)^2 / \text{tr } A'QA$$

Um die Symmetrie für A und B zu erreichen, wird ein zweites Maß erzeugt durch Multiplikation mit $u = \text{tr } B'QB / \text{tr } A'QA$

$$(3.3) \quad e_s = e u^{-1/2} = (\text{tr } A'QA)^{1/2} (\text{tr } B'QB)^{1/2} - \text{tr } (T'A'QB)^2 / (\text{tr } A'QA)^{1/2} (\text{tr } B'QB)^{1/2},$$

das die Symmetrie $e_s(A, B) = e_s(B, A)$ erfüllt.

Allerdings ist e_s von der Norm B abhängig. Um diesem Manko zu begegnen, reicht Division von e durch $\text{tr } B'B$:

$$(3.4) \quad L = \text{tr } E'E / \text{tr } B'B.$$

Ein weiteres Maß, nämlich

$$(3.5) \quad \text{tr } E'E / \text{tr } B'QB$$

erfüllt aber im Unterschied zu (3.4) neben der Skaleninvarianz auch das Symmetriekriterium. (3.5) läßt sich noch schreiben als

$$LS = \text{tr } E'E / \text{tr } B'QB = 1 - (\text{tr } T'A'QB)^2 / (\text{tr } A'QA \text{ tr } B'QB).$$

LS ist beschränkt durch 0 und 1. Wegen seiner Eigenschaften ist es den anderen Maßen vorzuziehen. Ferner kann man LS als Matrix-Analogon zu Guttman's coefficient of alienation (13) bezeichnen, da für den Fall $q = 1$ gilt

$$LS = 1 - \frac{(\sum_i (a_i - \frac{\sum a_i}{p}) (b_i - \frac{\sum b_i}{p}))^2}{\sum_i (a_i - \frac{\sum a_i}{p})^2 \sum_i (b_i - \frac{\sum b_i}{p})^2} = 1 - r^2(a,b).$$

Also ist LS das Matrix-Analogon zu $1 - r^2$, dem Bestimmtheitsmaß.

3.3 Ordinale Maße

Was über die Streßmaße Kruskals und Guttman's gesagt wurde, trifft auch auf die bisher betrachteten intervallskalierten Maße für die Ähnlichkeit von NMDS-Lösungen zu: sie messen auf anderem Niveau als auf dem

ordinalen Meßniveau der Daten und können daher besonders für Konfigurationen mit wenigen Punkten erhebliche Verzerrungen bewirken. Ordinale Maße der Relation zwischen $F(\delta_{ij})$ und $F(d_{ij})$, wie Spearman's rho, sind dagegen schon recht früh im Entwicklungsprozeß der MDS ausgesondert worden, weil sie "zu diskret für die Zwecke des Gradientenverfahrens sind" (Guttman, (13)).

$$(3.6) \text{ Spearman's rho: } \rho(a,b) = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

wobei d_i = Differenz der Rangplätze der einander entsprechenden Komponenten von a und b. $\rho \in [0,1]$.

Mit Korrektur für gebundene Werte:

$$(3.7) \rho_t(a,b) = \frac{T_a + T_b - \sum_i d_i^2}{2(T_a T_b)^{1/2}}$$

mit $T_a = 1/12 (n(n^2 - 1) - \sum_i t_{ai} (t_{ai}^2 - 1))$

$T_b = 1/12 (n(n^2 - 1) - \sum_i t_{bi} (t_{bi}^2 - 1))$

und $\sum_i t_{ai}, \sum_i t_{bi}$ = Anzahl der gebundenen Werte.

Kendall's τ

(3.8) Definitionen: Ein Permutationsschritt ist eine Vertauschung zweier benachbarter Komponenten einer Rangfolge. Eine andere Bezeichnung dafür ist "Inversion".

Beispiel: $F_1 = \langle a, b, c \rangle$

$F_2 = \langle a, c, b \rangle$

Folge F_2 entsteht durch Vertauschung von b mit c .

Seien $F_1 = \langle d_{1i} \rangle$ und $F_2 = \langle d_{2i} \rangle$ zwei Rangfolgen von Distanzen zwischen n Punkten, $i, j = 1, \dots, m$. Sei $S(F_1, F_2)$ die minimale Anzahl von Permutationsschritten, die notwendig ist, um F_1 in F_2 zu überführen. Dann heißt S "Permutationsabstand" zwischen F_1 und F_2 .

$s = S / m = S / n (n - 1)$ heißt normierter Permutationsabstand. $s \in [0, 1]$. Bildet man s ab auf $[-1, +1]$ durch $\tau = 2s - 1$, so ergibt sich Kendall's τ .

Oder anders: formt man die Information einer Rangfolge $\langle d_i \rangle$, $i = 1, \dots, m$ um zu

$$a_{ij} = \begin{cases} 1, & \text{falls } d_i < d_j \\ -1, & \text{falls } d_i > d_j \end{cases} \quad j = 1, \dots, n \quad i < j,$$

so läßt sich Kendall's τ -Koeffizient für zwei Rangfolgen $\langle d_{i1} \rangle, \langle d_{i2} \rangle$ schreiben als

$$\tau_{12} = \frac{\sum_{j>i} a_{ij1} a_{ij2}}{m(m-1)/2}$$

τ_{12} entspricht den Anforderungen des starken Monotoniekriteriums.

Die Korrekturform für gebundene Werte ist

$$\tau_{12} = \frac{\sum_{j>i} a'_{ij1} a'_{ij2}}{T_1 T_2}$$

$$T_1 = \left(\frac{m(m-1)}{2} - \sum_i \frac{t_{i1}(t_{i1}-1)}{2} \right)^{1/2}$$

$$T_2 = \left(\frac{m(m-1)}{2} - \sum_i \frac{t_{i2}(t_{i2}-1)}{2} \right)^{1/2}$$

t_{i1} bzw. t_{i2} sind die Anzahl der Ties in Folge $\langle d_{i1} \rangle$ und $\langle d_{j2} \rangle$.

$$a'_{ij} = \begin{cases} 1, & \text{falls } d_i > d_j \\ 0, & \text{falls } d_i = d_j \\ -1, & \text{falls } d_i < d_j \end{cases}$$

τ' entspricht den Anforderungen des abgeschwächten Monotoniekriteriums. τ' ist gleich τ , falls keine gebundenen Werte vorliegen. Beim schwachen Monotoniekriterium wird lediglich gefordert, daß aus $d_{i1} > d_{j1}$ folgt $d_{i2} \geq d_{j2}$. Ein angemessener, asymmetrischer Korrelationskoeffizient für diesen Fall wäre

$$(3.10) \quad \tau_{12}^* = \frac{\sum_{j > i} S_{ij}^*(12)}{\frac{m(m-1)}{2}}$$

$$\text{mit } S_{ij}^*(12) = \begin{cases} 1, & \text{falls } a_{ij1} \cdot a_{ij2} \geq 0 \\ -1 & \text{sonst} \end{cases}$$

Für nichtgebundene Daten sind alle obigen Varianten von τ identisch, während für gebundene Daten gilt, daß $\tau^* \geq \tau' \geq \tau$ ist.

Guttman's Monotonie-Koeffizient für schwache Monotonie läßt sich schreiben als

$$\mu_0(12) = \frac{\sum_{j>i} a'_{ij1} a'_{ij2}}{\sum_{j>i} |a'_{ij1} a'_{ij2}|}$$

Er kommt daher dem τ' nahe, hat jedoch einen größeren Zähler als τ' , falls Ties in beiden Folgen übereinstimmen und einen ganzzahligen Nenner, während der Nenner von τ' reell ist. Der Nenner von μ_0 ist im allgemeinen größer als der von τ' (26).

Da die τ -Maße und ρ keinerlei Bedingungen wie Intervallskaliertheit der Meßwerte oder lineare Abhängigkeit aufstellen, sind sie in idealer Weise geeignet, Ähnlichkeiten zwischen MDS-Konfigurationen und zwischen Distanz- und Unähnlichkeitsmatritzen zu berechnen - zumindest dann, wenn nur die Ordnungsrelationen bekannt sind und kein inhaltliches Vorwissen gleichfalls berücksichtigt werden muß.

3.4 Beziehungen zwischen metrischen und ordinalen Korrelationsmaßen

Da wir es in der NMDS sowohl mit metrischen als auch ordinalen Größen zu tun haben und einen Vergleich zwischen ihnen herstellen wollen, müßte an dieser Stelle ein Maß für den Zusammenhang (Korrelation) von ordinalen und intervallskalierten Variablen erwähnt werden. Ein solches Maß existiert jedoch nicht. Wir wollen stattdessen Korrelationskoeffizienten für metrische mit denen für ordinale Variablen vergleichen.

Ausgangspunkt jedes Korrelationskoeffizienten ist ein Maß für die Streuung beider Variablen sowie für das gemeinsame Variieren. Als gebräuchliche Streuungsmaße kommen in Betracht:

a) die Variationsweite (range), $v = x_{\max} - x_{\min}$

b) der Interquartilbereich oder Interdezilbereich
 $v = Q_1 - Q_3$ oder $v = D_9 - D_1$

c) die durchschnittliche Abweichung $v = \frac{1}{n} \sum_i f_i |x_i - \bar{x}|$

d) die Varianz

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

f_i = Häufigkeit des i-ten Meßwerts

$\bar{x}$ = arithmetisches Mittel

Q_i = i-tes Quartil

D_i = i-tes Dezil

Während Variationsweite und Interquartilbereiche wegen ihrer Instabilität bei Ausreißern sowie ihrer schlechten Schätzeigenschaften für die Streuung in der Gesamtpopulation meist nur für den groben Überblick bei Schätzproblemen benutzt werden, ist c) suffizienter Schätzer und d) effizient und erwartungstreu, falls x normalverteilt ist. Der besondere Zusammenhang von d) mit der Normalverteilung sowie die einfachere algebraische Handhabung haben d) und nicht c) zum "Standard"-Abweichungsmaß werden lassen.

Als Maß des gemeinsamen Variierens zweier intervallskalierter Variablen wird nahezu ausschließlich die Kovarianz $\text{cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$ genommen.

In einigen Fällen könnte man sicher ebensogut

$\text{range}(x_i - y_i) = \max_i (x_i - y_i) - \min_i (x_i - y_i)$ neh-

men, falls die Skala für x und y normiert ist, und

die Beiträge der Einzelwerte unerheblich sind.

Ebenso ist $\sum_i (x_i - y_i) (\theta_x - \theta_y)$ plausibel (θ_x, θ_y seien Mittelwerte), falls die Differenz von x und y einen Sinn ergibt. Das ist dann der Fall, wenn x und y den gleichen Wertebereich haben.

Es ergeben sich folgende Möglichkeiten der Definition eines Korrelationskoeffizienten:

$$(3.11) \quad r = \frac{n-1}{n} \frac{\sum_i (x_i - \theta_x) (y_i - \theta_y)}{\sqrt{\sum_i (x_i - \theta_x)^2 \sum_i (y_i - \theta_y)^2}}$$

$$\text{Im Falle } \theta = \text{arithmetisches Mittel gilt } r = \frac{\text{cov}(x, y)}{(\text{var}(x) \text{var}(y))^{1/2}}$$

(Pearson'scher Maßkorrelationskoeffizient)

Dieser Schätzer knüpft an die guten Eigenschaften der Kovarianz und Varianz an. Er setzt die Linearitätshypothese voraus. Tatsächlich ist er nur dann ein verlässlicher Schätzer für den Zusammenhang zweier Variablen, wenn sich dieser als linearer Zusammenhang ergibt. Bei funktionalen Zusammenhängen der Art $y = x^2$, $x \in [-2, 2]$ wird $r = 0$, d.h. r läßt fälschlich schließen auf die Unabhängigkeit von x und y . Es soll daher eine weitere Klasse von Korrelationskoeffizienten dargestellt werden. Hierbei wird zur Voraussetzung gemacht, daß x_i und y_i direkt vergleichbar sind, d.h. der Wert $(x_i - y_i)$ eine sinnvolle Größe darstellt. Dies kann erreicht werden z.B. durch Normierung.

$$(3.12) \quad r = \frac{\sum_i w_{xy} (\theta_x - \theta_y) (x_i - y_i)}{\sum_i w_{xy} |\theta_x - \theta_y| |x_i - y_i|}$$

für $\sum_i w_{xy} |\theta_x - \theta_y| |x_i - y_i| \neq 0$;

w_{xy} = nichtnegative Gewichte

$$\begin{aligned} \text{Wegen } -|\theta_x - \theta_y| |x_i - y_i| &\leq (\theta_x - \theta_y)(x_i - y_i) \\ &\leq |\theta_x - \theta_y| |x_i - y_i| \end{aligned}$$

gilt $-1 \leq r \leq +1$.

r kann leicht erweitert werden für den Fall mehrerer Populationen P_k , $k = 1, \dots, l$ mit Zufallsvariablen X_k , $k = 1, \dots, l$:

Der Zähler von r wird dann zu $\sum_i w_{k_1 k_2 i} (\theta_{k_1} - \theta_{k_2})(x_i - y_i)$,
der Nenner zu $\sum_i w_{k_1 k_2 i} |\theta_{k_1} - \theta_{k_2}| |x_i - y_i|$

Es läßt sich dann ein neuer Koeffizient erzeugen durch

$$(3.13)r = \frac{\sum_i \sum_{k_1=1}^l \sum_{k_2=1}^l w_{k_1 k_2 i} (\theta_{k_1} - \theta_{k_2})(x_i - y_i)}{\sum_i \sum_{k_1} \sum_{k_2} w_{k_1 k_2 i} |\theta_{k_1} - \theta_{k_2}| |x_i - y_i|}$$

$$x_i \in P_{k_1} ; y_i \in P_{k_2}$$

Auch hier gilt wieder $-1 \leq r \leq 1$.

Durch spezielle Wahl von $w_{k_1 k_2 i}$ in (3.13) bzw. w_{xy}

in (3.12) lassen sich spezielle Koeffizienten erzeugen. Setzt man z.B.

$$w_{k_1 k_2 i} = \frac{x_i - y_i}{\theta_{k_1} - \theta_{k_2}} \quad \text{und ist } \theta \text{ das arithmetische Mit-}$$

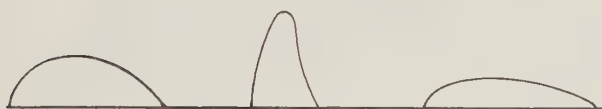
tel, so ergibt sich aus (3.13)

$$(3.14) \quad r = \frac{\sum_{k_1} \sum_{k_2} \sum_i \text{sign}(\theta_{k_1} - \theta_{k_2}) \cdot \text{sign}(x_i - y_i) \cdot (x_i - y_i)^2}{\sum_{k_1} \sum_{k_2} \sum_i (x_i - y_i)^2}$$

Dieser Koeffizient wird gleich 1, falls $\text{sign}(x_i - y_i)$ und $\text{sign}(\theta_{k_1} - \theta_{k_2})$ in allen Fällen übereinstimmen, d.h. falls $\theta_{k_1} - \theta_{k_2} > 0$, so auch $x_i - y_i > 0$ ist und damit die Isotonie $\theta_{k_1} > \theta_{k_2} \rightarrow x_i > y_i$ gewahrt bleibt. $r = -1$, falls die Isotonie in jedem Fall verletzt wird, also die Rangfolge der θ_k und der dazugehörigen x_i, y_i invertiert ist. Im Vergleich zu Pearson's Maßkorrelation mißt (3.14) nicht, ob x_i und y_i in gleicher Richtung vom Mittelwert abweichen, sondern ob die Rangfolge, die die Mittelwerte vorgeben, von den Realisationen beibehalten wird. Falls x und y in der Weise verteilt sind, daß die Dichtefunktion eingipflige Kurven sind, gilt, daß der Koeffizient (3.11) stets kleiner gleich demjenigen aus (3.14) ist.

(3.14) wird gleich eins, falls sich die empirischen Verteilungen von P_{k_1} nicht überlappen. Weiterhin ist (3.14) nicht an die Linearität der Beziehung zwischen x und y gebunden. Bei zwei Populationen P_1 und P_2 ergeben alle diejenigen Differenzen $(x_i - y_i)$ einen positiven Beitrag zu (3.14), die unterhalb der Diagonalen $y = x$ liegen (bei $\theta_1 - \theta_2 > 0$), im dreidimensionalen diejenigen im Unterraum $\left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mid x < y < z \right\}$ falls $\theta_1 < \theta_2 < \theta_3$ etc. (15).

Guttman (15) benutzte einen etwas veränderten Koeffizienten zur Messung des Grades der Diskriminanz durch Verteilungen auf einer Achse. Sind die Verteilungen nicht überlappend, so ist es möglich, die Zugehörigkeit eines gemessenen Wertes zu einer Population exakt zu bestimmen:



Wählt man die Schnittpunkte x' und x'' wie folgt:



so ist die vollständige Diskriminanz gewährleistet. Guttman wählte für diese Situation zwei Koeffizienten, die er "coefficients of efficacy" nennt:

$$(3.15) \mu_2 = \frac{\sum_{k=1}^m (\theta_k - \theta)^2}{\frac{1}{2N} \sum_{k_1} \sum_{k_2} |\theta_{k_1} - \theta_{k_2}| \sum_{i \in P_{k_1}} \sum_{j \in P_{k_2}} |x_i - y_j|}$$

und

$$(3.16) \mu_1 = \frac{\sum_{k_1} \sum_{k_2} |\theta_{k_1} - \theta_{k_2}|}{\sum_{k_1} \sum_{k_2} (1 - \delta_{k_1 k_2}) \sum_{i \in P_{k_1}} \sum_{j \in P_{k_2}} |x_i - y_j|}$$

$$\delta_{k_1 k_2} = \begin{cases} 1 & \text{falls } k_1 = k_2 \\ 0 & \text{sonst} \end{cases}$$

Man beachte, daß hier über zwei Indizes $i \in P_{k_1}$ und $j \in P_{k_2}$ getrennt summiert wird, so daß jeder Wert aus einer Populationsverteilung mit jedem der Werte aus den anderen verglichen wird. μ_2 läßt sich wiederum durch den Pearson'schen Korrelationsquotienten η abschätzen:

η variiert zwischen 0 und 1 und läßt sich schreiben als

$$(3.17) \quad \eta = 1 - \sqrt{\frac{\sum \sigma_k^2 N_k}{N \sigma^2}}$$

wobei σ_k^2 und N_k Varianz und Gewicht der k -ten Population sind, σ^2 und N Varianz und Gewicht der Gesamtpopulation.

$\eta = 0$ genau dann, wenn für jedes k ^(gilt) $\sqrt{\theta_k} = \theta$, θ_k arithmetisches Mittel der k -ten Population, θ arithmetisches Mittel der Gesamtpopulation. Das kann leicht durch die Umformung

$$\eta = \sqrt{\frac{\sum_k N_k (\theta_k - \theta)^2}{N \sigma^2}} \quad ,$$

$$N \sigma^2 = \sum_k N_k \sigma_k^2 + \sum_k N_k (\theta_k - \theta)^2$$

gezeigt werden, d.h. $N \sigma^2$ wird aufgespalten in die Sum-

me der Varianz innerhalb und der zwischen den Populationen. $\eta = 1$ falls die Summe der Varianzen innerhalb der Populationen $\sum_k N_k \sigma_k^2$ verschwindet, was aber nur selten der Fall sein dürfte. Ferner ist zu bemerken, daß η auch für das Problem der Diskriminanz zwischen Populationen kein geeignetes Maß ist - η kann selbst bei vollständig möglicher Diskriminanz beliebig nahe bei Null liegen, wählt man nur $\sum_k N_k \sigma_k^2$ groß genug.

Zum Vergleich: falls $\eta = 0$ oder 1 gilt $\mu_2 > \eta$, da für den Nenner von μ_2 die Schwarz'sche Ungleichung gilt: Nenner von μ_2

$$\leq \sqrt{\left[\sum_{k_1} \sum_{k_2} \sum_i \sum_j (\theta_{k_1} - \theta_{k_2})^2 \right] \cdot \left[\sum_{k_1} \sum_{k_2} \sum_i \sum_j (x_i - y_i)^2 \right]}$$

$$= \sqrt{4 N^2 \sigma^2 \sum_k N_k (\theta_k - \theta)^2}$$

Setzt man den letzten Term als Nenner in μ_2 ein, so erhält man η . Also $\mu_2 \geq \eta$ und $\mu_2 = \eta$ falls $\eta = 1$ oder $\eta = 0$. D.h. $\eta \leq \mu_2 \leq 1$.

μ_2 und μ_1 sind gleich 0, falls alle arithmetischen Mittel θ_k identisch sind, d.h. keine Diskriminanz möglich ist. Der Wert 1 wird angenommen, wenn sich die Populationen nicht überlappen. Werte zwischen 0 und 1 geben den Grad der durchschnittlichen Diskriminanzmöglichkeiten an. Eine allgemeine Form analog zu (3.13) kann angegeben werden mit

$$(3.18) \quad \mu = \frac{\sum_{k_1} \sum_{k_2} \sum_i \sum_j w_{k_1 k_2 i j} (\theta_{k_1} - \theta_{k_2})(x_i - y_i)}{\sum_{k_1} \sum_{k_2} \sum_i \sum_j w_{k_1 k_2 i j} |\theta_{k_1} - \theta_{k_2}| |x_i - y_i|}$$

Bezeichnungen analog zu denen von (3.13).

Anwendung bei der MDS

Die hier vorgestellten Korrelationskoeffizienten sind ausgezeichnet geeignet, den Zusammenhang zwischen zwei Rangfolgen von Distanzen zu messen, sowie auch der Monotonieanforderung zwischen Distanzen und Unähnlichkeiten gerecht zu werden. Denn hierbei ist nicht die Linearität eine geeignete Oberhypothese - wie sie etwa in dem Testverfahren von Lingoes / Borg angegeben wurde, sondern die Isotonie.

Es soll gezeigt werden, daß unter gewissen Voraussetzungen Guttman's coefficient of efficacy in engem Zusammenhang zu Kendall's τ^u steht.

Analog zu den Populationen P_{ki} betrachten wir hier Stichprobenräume, und zwar für jeden der Rangplätze $k'=1, \dots, m$ einer Rangfolge von Distanzen das Intervall $(0, \infty)$ als Wertebereich der möglichen Realisationen von $d_{i_k, j_{k'}}$. Nach geeigneter Transformation der d_{ij} und der zugehörigen δ_{ij} kann angenommen werden, daß jedes der d_{ij} einer Verteilung mit dem Parameter (einer zentralen Tendenz) δ_{ij} entstammt. Darüber ist die Abhängigkeit zwischen den δ_{ij} und d_{ij} vermittelt. Annahmen über die Art der Verteilung der d_{ij} bzw. über die Unabhängigkeit der d_{ij} voneinander werden hier noch nicht gemacht.

Ein geeignetes Maß für die Korrelation der d_{ij} mit den dazugehörigen δ_{ij} ist

$$(3.19) \quad r_d = \frac{\sum_{k_1=1}^m \sum_{k_2=1}^m (\theta_{k_1} - \theta_{k_2}) (d_{k_1} - d_{k_2})}{\sum_{k_1=1}^m \sum_{k_2=1}^m |\theta_{k_1} - \theta_{k_2}| |d_{k_1} - d_{k_2}|}$$

eine Abwandlung von (3.18) mit $w_{k_1 k_2 i j} = 1$.

Die oben erwähnte Transformation der d_{ij} und δ_{ij} werde so durchgeführt, daß

$$(3.20) \quad \delta_{i_1 j_1} = \frac{1}{m+1}, \quad \delta_{i_m j_m} = \frac{m}{m+1}, \quad (\delta_{i_1 j_1} - \delta_{i_{l+1} j_{l+1}}) = \frac{1}{m+1}$$

sowie $\max_{d_{i_k' j_k'}} |d_{i_k' j_k'}| \leq 1$.

Das ist möglich, weil die Distanzen der NMDS translationsinvariant und die δ ordinal sind.

Verzichtet man ferner völlig auf den Beitrag der Differenzen zwischen Rangplätzen und berücksichtigt nur die Verletzung der Isotonie, so kann man in (3.18) setzen

$$w_{k_1 k_2 i j} = \frac{1}{|\delta_{k_1} - \delta_{k_2}| |d_{k_1} - d_{k_2}|}$$

und erhält

$$(3.21) \quad r_d = \frac{\sum_{k_1} \sum_{k_2} \text{sign}(\theta_{k_1} - \theta_{k_2}) \text{sign}(d_{k_1} - d_{k_2})}{m(m-1)} =$$

$$\sum_{k_1} \sum_{k_2} \frac{a_{k_1 k_2}^1 \cdot a_{k_1 k_2}^2}{m(m-1)} = \text{Kendall's tau},$$

die $a_{k_1 k_2}$ sind erklärt wie in (3.9).

ein
Dadurch ist Zusammenhang zwischen Guttman's coefficient of efficacy und Kendall's tau gezeigt.

Ein weiterer Zusammenhang ergibt sich über den Erwartungswert von r_d aus (3.19). Der Beweis dazu setzt allerdings einschneidende Verteilungsannahmen voraus. Es seien zuvor einige Ausführungen von Lingoes zitiert, der Ähnliches über den Zusammenhang von Pearson's r und Spearman's ρ formulierte.

Lingoes (1980) stellte im Zusammenhang mit der Diskussion um die Unabhängigkeitsvoraussetzung für die abhängige Variable Y beim Hotelling'schen t -Test eine Beziehung zwischen metrischen und ordinalen Korrelationskoeffizienten her. Zunächst stellte er zwei Randomisierungsannahmen für Y und X auf, nachdem er die Robustheit des t -Tests bezüglich der Verletzung der Normalverteiltheit gezeigt hatte:

- 1) Seien $x_1, \dots, x_k$ und y Rangbilder von Zufallsvariablen, die im m -dimensionalen Intervall $[0, 1]^m$ uniform-verteilt sind. Damit wird zugleich sichergestellt, daß alle Variablen identisch und unabhängig voneinander verteilt sind, es keine Abhängigkeiten innerhalb der Mengen (der $\{x_i\}$) gibt und daß es gute Chancen für lineare Abhängigkeit (zwischen x und y) gibt.
- 2) Prädiktor und abhängige Variable seien eine Menge von Rangnummern von 1 bis m , wobei die Rangnummern in derselben Anordnung stehen wie die entsprechenden Komponenten der Variablen. Die abhängige Variable sei eine zufällige Permutation und ist eine Stichprobe aus einer Population von gleichwahrscheinlichen Permutationen. Diese Rangnummern stellen wie in 1) sicher, daß keine Abhängigkeiten innerhalb der Mengen auftreten und ermöglichen die Linearität, vorausgesetzt, daß n hinreichend groß ist ($n \geq 10$), so daß die Werte sich einem kontinuierlichen Maß annähern (vgl. hierzu auch (18)).

Um zu zeigen, daß beide Annahmen äquivalent sind, führt er Simulationsergebnisse an, nach denen die Verteilung von t unter beiden Annahmen gleich ist und einen Satz, nach dem die Spearman'sche Rangordnungskorrelation gleich dem Erwartungswert der regulären Produktmomentkorrelation ist, wenn X und Y wie

in 1) sind. Ferner sagt er, daß "... wenn wir annehmen, daß unsere Skalen rein ordinal sind und die ranggeordneten X und Y uniformverteilt, das einzig beste Maß für die lineare Beziehung Spearman's rho ist und nicht Pearson's r."

Die Uniformverteilung erscheint aber für Distanzen der MDS nicht adäquat zu sein. Aus empirischen Erwägungen könnte man für die d_i^2 eine χ^2 -Ver-

teilung annehmen. (vgl. (52) S. 56) Die Uniformverteilung berücksichtigt nicht die Abhängigkeiten zwischen den Distanzen. Trotzdem soll noch eine ähnliche Aussage für Kendall's tau bewiesen werden: Man führe o.B.d.A. die obige Transformation (3.20) der d_k und δ_k in das Intervall $[0,1]$ durch. Dann gilt:

Satz

Setzt man voraus, daß jedes der d_k , über dem Intervall $[0,1]$ uniform verteilt ist und betrachtet nur den m-dimensionalen Teilbereich, in dem sich die $\{d_k\}$ monoton zueinander verhalten, d.h. die gleiche Rangordnung $F(d_k): \hat{=} d_1, \leq d_2, \leq \dots \leq d_m$, besitzen, so gilt:

Der Erwartungswert des Korrelationskoeffizienten r_d ,

$$E(r_d) = \frac{\sum_{k_1} \sum_{k_2} \text{sign}(\theta_{k_1} - \theta_{k_2}) (d_{k_1} - d_{k_2}) / (k'_1 - k'_2)}{\sum_{k_1} \sum_{k_2} (d_{k_1} - d_{k_2}) / (k'_1 - k'_2)}$$

$$k'_1 = E(d_{k_1})$$

ist gleich Kendall's Rangkorrelationskoeffizient τ .

Beweis:

Nach David (18), S. 27) ist der Erwartungswert $\mu_{r:n}$ des r -ten Elements einer Rangfolge von n geordneten, uniformverteilten Zufallsvariablen gleich $r / m + 1$, wobei der Wertebereich in $(m + 1)$ - gleichgroße Abschnitte unterteilt wird. Also gilt hier

$$E(r_d) = E \frac{\gamma \sum_{k_1} \sum_{k_2} \frac{d_{k_1} - d_{k_2}}{k'_{11} - k'_{12}} - (1 - \gamma) \sum_{k_1} \sum_{k_2} \frac{d_{k_1} - d_{k_2}}{k'_{11} - k'_{12}}}{\gamma \sum_{k_1} \sum_{k_2} \frac{d_{k_1} - d_{k_2}}{k'_{11} - k'_{21}} + (1 - \gamma) \sum_{k_1} \sum_{k_2} \frac{d_{k_1} - d_{k_2}}{k'_{11} - k'_{21}}}$$

$$\gamma = \begin{cases} 1 & \text{falls } \text{sign}(\delta_{k_1} - \delta_{k_2}) = \text{sign}(d_{k_1} - d_{k_2}) \\ 0 & \text{sonst} \end{cases}$$

$$= \frac{\gamma \sum_{k_1} \sum_{k_2} \frac{k'_{11} - k'_{21}}{k'_{11} - k'_{21}} - (1 - \gamma) \sum_{k_1} \sum_{k_2} \frac{k'_{11} - k'_{21}}{k'_{11} - k'_{21}}}{m(m-1)}$$

$$= \frac{\sum_{k_1} \sum_{k_2} a_{k_1 k_2 1} a_{k_1 k_2 2}}{m(m-1)} = \text{Kendall's } \tau.$$

Man könnte nun versucht sein, Beziehungen zwischen metrischen und ordinalen Korrelationskoeffizienten über die Bildung von Erwartungswerten zu suchen. Falls die Verteilung der metrischen Variablen bekannt ist, ist das möglicherweise auch von Nutzen. Bei der MDS ist die metrische Variable nun gerade die zu bestimmende, deren Verteilung unbekannt ist. Die Verteilungsannahme ist aber entscheidend für das Schätzergebnis des Erwartungswertes. Also dürfte diese Form

der Bezugnahme nicht stattfinden, d.h. es ist nicht möglich, zu sagen : "Wenn die Daten rein ordinal sind, so ist das beste Maß für die Erfüllung der Isotoniebeziehung Kendall's τ als Erwartungswert von r_d ." Eine solche Aussage setzt vielmehr die Verteilungsannahme voraus. Aus dem obigen Satz ergibt sich aber eine Motivation zur Benutzung von Kendall's τ als Maß für die Verletzung der Isotonieanforderung.

4. Kapitel Hypothesentests auf Mengen von Rangfolgen

4.1 Testverfahren

Ziel dieser Arbeit ist, ein Testverfahren für NMDS-Darstellungen zu entwickeln, das sich nur auf die ordinalen Informationen des Datenmaterials stützt und diese mit dem in der NMDS-Lösung enthaltenen ordinalen Informationen vergleicht. Ein solches Testverfahren hat seine Bedeutung in der Möglichkeit, mit der Bestimmung der Fehler erster und zweiter Art Aussagen über die Adäquatheit der NMDS-Lösung für die zugrundeliegenden Daten zu machen. Die bisher übliche ausschließliche Benutzung des Streßmaßes als Kriterium für den Grad der Angepaßtheit eines Modells findet ihre Grenzen vor allem darin, daß

1. keine statistischen Kriterien für die Ablehnung eines Modells existieren und
2. keine Aussage über die fälschliche Akzeptierung eines Modells gemacht werden kann.

Wie wir in Kapitel 3 gesehen haben, gilt Punkt 2. auch für die vorgeschlagenen "metrischen" Testverfahren. Im Unterschied zu herkömmlichen Hypothesentests ist beim Testen von Modellen aber gerade dieser Fehler zweiter Art von besonderer Bedeutung.

Es geht in der Inferenzstatistik meist gerade darum, sich nicht fälschlicherweise gegen eine kritische Hypothese H_0 zu entscheiden, wie z.B. "die Gruppen der Deutschen und Ausländer verhalten sich bezüglich der Anzahl von Unfällen am Arbeitsplatz gleich" oder "eine Behandlung mit Placebos hat den gleichen Effekt wie die mit einem Medikament", da eine Ablehnung sol-

cher Hypothesen erhebliche Konsequenzen für sozialwissenschaftliche Theoriebildung und / oder Änderung von praktischen Handlungsmaximen haben kann. Man formuliert daher die kritische Hypothese H_0 als diejenige, deren fälschliche Ablehnung schwerwiegende Folgen haben kann, und legt das Entscheidungsverfahren so fest, daß der Fehler erster Art möglichst unwahrscheinlich wird, etwa mit der Fehlerwahrscheinlichkeit (zum Signifikanzniveau) von 5% oder 1%. Der Fehler zweiter Art, d.h. die fälschliche Ablehnung der "Gegenhypothese" H_1 , kann darüber oft vernachlässigt werden oder wird nur als Optimalitätskriterium bei der Auswahl von statistischen Testverfahren benutzt (sog. trennschärfste Tests).

Im Fall der Modellschätzung liegt das Problem der Fehlerbestimmung erster und zweiter Art anders. Wegen der Komplexität von Modellen kann die H_0 -Hypothese fast immer nur heißen: "Das Modell ist mit den Daten kompatibel", da nur unter dieser Hypothese eine Verteilung für eine statistische Prüfgröße überhaupt bestimmt werden kann. Würde man als kritische Hypothese aufstellen "ein anderes Modell als das vorliegende ist mit den Daten kompatibel", so müßte die Verteilung einer statistischen Prüfgröße über sämtliche Alternativmodelle ermittelt werden, ein Problem, das fast immer weder berechenbar noch überhaupt sinnvoll mathematisch formulierbar ist. Obwohl also diese Hypothese die für Modellannahmen eigentlich kritische ist, die es zu widerlegen gilt, kann man meist nur die andere prüfen, die die Angepaßtheit eines bestimmten Modells zum Inhalt hat. Aus demselben Grund kann auch der Fehler zweiter Art nicht bestimmt werden, da dessen Wahrscheinlichkeitsverteilung von der Gegenhypothese und damit den Ausprägungen aller anderen denkbaren Modelle abhängt.

Ein Hypothesenprüfen im herkömmlichen statistischen Sinn kann daher für Modelle meist nicht geleistet werden: Alternativmodelle werden im Testverfahren überhaupt nicht berücksichtigt, es sei denn, sie seien aufgrund inhaltlicher Überlegungen bereits bekannt und könnten selbst als H_0 -Hypothesen formuliert bzw. gegen ein anderes Modell geprüft werden.

Die Verfahren der NMDS "schätzen" metrische Abstände zwischen Punkten im R^D aufgrund ordinaler Daten, d.h. einer Rangfolge von Unähnlichkeiten zwischen Objekten, die als Punkte dargestellt werden. Wie in Kapitel 2 dargestellt, ist das Ziel einer Skalierung, ein isotones Verhältnis zwischen den metrischen Distanzen und den Unähnlichkeiten herzustellen. Die Distanzen werden also so bestimmt, daß ihre Rangfolge möglichst ähnlich derjenigen der Unähnlichkeiten ist.

Da die Ausgangsdaten ordinaler Natur sind und auch das Ziel der NMDS nach ordinalen Gesichtspunkten - Ähnlichkeit von zwei Rangfolgen - beurteilt wird, muß ein adäquates Prüfmaß ebenfalls das ordinale Niveau berücksichtigen. Für die Fragestellung, ob eine NMDS-Darstellung ihren ordinalen Ausgangsdaten überhaupt angemessen ist, darf daher einer solchen Darstellung nicht ihr metrischer Charakter zugute kommen, wie es beispielsweise im Streßmaß geschieht.

Das ordinale Testverfahren soll nun prüfen, ob für gegebene Ausgangsdaten eine NMDS-Konfiguration adäquat ist. Es ist ordinal insofern, als es untersuchen soll, ob die ordinalen Informationen von NMDS-Konfigurationen, also die Rangfolgen ihrer Distanzen, in ausreichender Weise die ordinalen Informationen der Ausgangsdaten wiedergeben. Ferner soll die Wahrscheinlichkeit des Fehlers zweiter Art bestimmt werden können, beru-

hend auf der Gegenhypothese: "Den Ausgangsdaten entspricht keine Darstellung im R^P ".

4.2 Definitionen

Ein plausibles Maß für den Abstand zwischen zwei Rangfolgen oder Permutationen ist "S", d.i. die kleinste Anzahl der paarweisen Vertauschungen zwischen zwei benachbarten Komponenten, die notwendig ist, um eine Rangfolge in die andere zu permutieren. Die drei Folgen

$$F_1 = \langle a, b, c, d, e \rangle$$

$$F_2 = \langle c, b, a, d, e \rangle$$

$$F_3 = \langle e, d, c, b, a \rangle$$

haben die Permutationsabstände

$$S(F_1, F_2) = 3$$

$$S(F_1, F_3) = 10$$

$$S(F_2, F_3) = 7$$

S ist zudem eng verwandt mit Kendall's tau-Maß, das wiederum Ähnlichkeiten mit Guttman's Alienationskoeffizienten aufweist (vgl. Kapitel 3.3).

Ist n die Anzahl der Objekte, zwischen denen Unähnlichkeiten gemessen wurden, so ist $m = n(n - 1) / 2$ die Anzahl der Unähnlichkeiten und ebenfalls die Anzahl der Komponenten der zu betrachtenden Rangfolge.

Der maximale Abstand S zwischen zwei Rangfolgen mit m Komponenten ist $\max S = m(m - 1) / 2$.

Die Zahl der verschiedenen Rangfolgen mit m Elementen ist $m!$. Das bedeutet, daß die Anzahl der verschiedenen Rangfolgen schon bei wenigen Punkten astronomisch groß ist:

<u>Anzahl der Punkte</u>	<u>Anzahl der Un-</u>	<u>Anzahl der Rang-</u>
	<u>ähnlichkeiten</u>	<u>folgen</u>

1.....	1.....	1
2.....	2.....	2
3.....	3.....	6
4.....	6.....	720
5.....	10.....	$3,62880 \cdot 10^7$
6.....	15.....	$1,30767 \cdot 10^{13}$
7.....	21.....	$5,10909 \cdot 10^{20}$
8.....	28.....	$3,04888 \cdot 10^{29}$
9.....	36.....	$3,71993 \cdot 10^{41}$
10.....	45.....	$1,19622 \cdot 10^{56}$
11.....	55.....	$1,26964 \cdot 10^{73}$
12.....	66.....	$5,44344 \cdot 10^{92}$

Von den möglichen Rangfolgen für feste Punktzahlen unterscheiden sich einige nur durch die Namensgebung der Punkte, d.h. vertauscht man die Punktenamen, so kann man dadurch eine Rangfolge in die andere über-

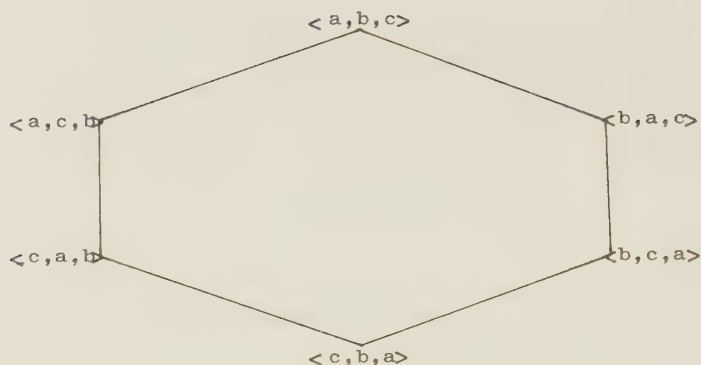
führen. Man kann alle Rangfolgen, die bis auf Vertauschung von Punktenamen identisch sind, zu Klassen zusammenfassen. Die Repräsentanten dieser Klassen sollen als "wesentliche Rangfolgen" bezeichnet werden.

Ihre Anzahl ist $\frac{m!}{n!}$. Der Repräsentant zu einer Folge F heißt $r(F)$.

Ferner sollen noch diejenigen Rangfolgen, die ohne Verletzung des Isotoniekriteriums in eine räumliche Konfiguration im R^p überführt werden können, "im R^p darstellbare Rangfolgen" oder einfach nur "darstellbare Rangfolgen" heißen. Darstellbare Rangfolgen sind identisch mit Rangfolgen von Distanzen. Ihr Streß ist $= 0.0$. Wie bereits erwähnt (Kapitel 2), ist bei genügend großer Dimension p des Darstellungsraums jede Rangfolge darstellbar. Der Begriff der Darstellbarkeit ist daher nur in Zusammenhang mit einer bestimmten Dimension sinnvoll. Die Elemente einer Klasse von darstellbaren Rangfolgen haben geometrisch kongruente Konfigurationen. Die Menge der darstellbaren Rangfolgen wird bezeichnet mit "D".

4.3 Darstellung der Rangfolgen als Graph, Beispiele

Die Rangfolgen mit fester Komponentenzahl m können in einem Graphen zusammen dargestellt werden, wobei die einzelnen Rangfolgen die Ecken darstellen und jede der Kanten einen Permutationsschritt bedeutet, durch den eine Rangfolge in ihren Nachbarn überführt wird. Für $n = 3$; $m = 3$ ergibt sich ein sechseckiges Gebilde mit



In jede Ecke münden $m - 1$ Kanten, da $m - 1$ einfache Vertauschungen in jeder Rangfolge durchgeführt werden können. Für $n > 3$ ist jede Ecke gleichzeitig Ecke von sechseckigen und viereckigen Gebilden. Die viereckigen Gebilde *) entsprechen zyklischen Vertauschungen von Komponenten der Rangfolge, bei denen nicht zweimal nacheinander die gleiche Komponente betroffen ist, z.B.

$\langle a, b, c, d \rangle$ $\langle b, a, c, d \rangle$
 $\langle a, b, d, c \rangle$ $\langle b, a, d, c \rangle$

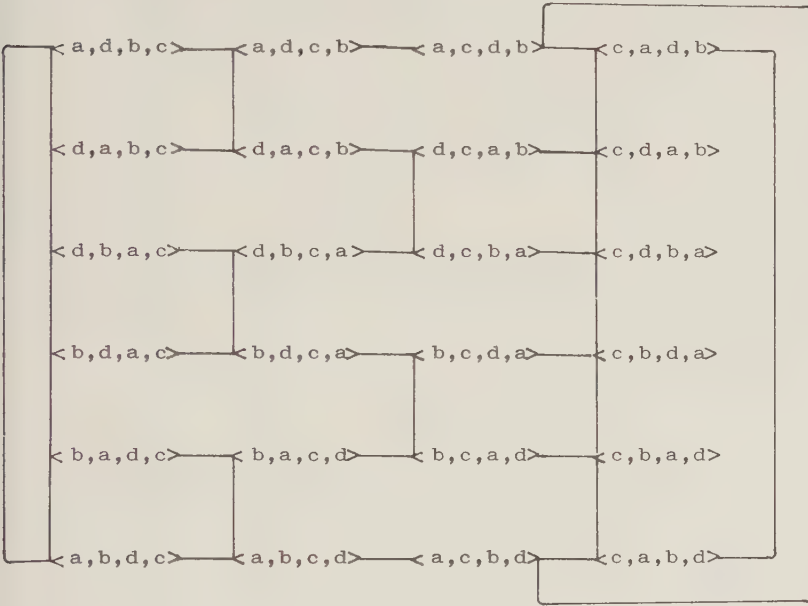
In den Sechsecken wird zweimal nacheinander die selbe Komponente vertauscht. Jede Ecke liegt in

$$\sum_{i=1}^{m-3} I = (m - 2)(m - 3) / 2$$

Vierecken und $(m - 2)$ Sechsecken.

*) Die "Gebilde" zeichnen sich dadurch aus, daß die kürzeste Verbindung zwischen zwei ihrer Ecken über ihre Kanten zu suchen ist.

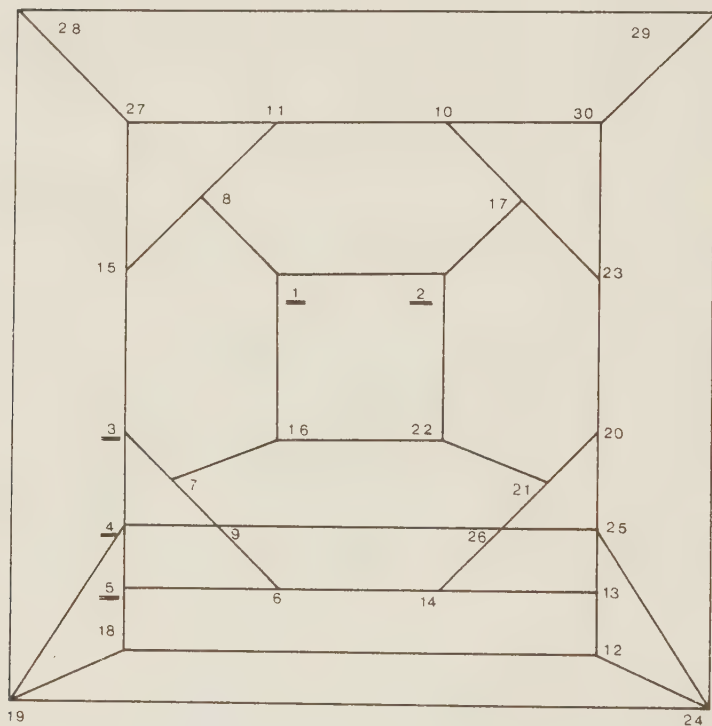
Beispiel: $m = 4$ (dem entspricht keine Zahl n von Objekten):



Der Fall $n = 4$, $m = 6$ mit 720 Ecken ist, wie leicht nachzuprüfen, nicht mehr in der Ebene darstellbar, ohne daß sich Kanten überschneiden. Zur Darstellung des Graphen kann hier und für alle größeren n der R^3 dienen (vgl. (32)).

Es seien daher hier nur die wesentlichen Rangfolgen für $n = 4$ mit Punkten A, B, C, D wiedergegeben. Als Repräsentanten wurden die Rangfolgen ausgewählt, in denen $a = S(A, B)$ der kleinste Abstand ist; ferner soll $S(A, C) = b$ oder $S(C, D) = f$ der zweitkleinste Abstand sein und, falls f der zweitkleinste, so ist b der drittkleinste Abstand. Dadurch ist die Namensgebung eindeutig festgelegt.

Graph der wesentlichen Rangfolgen von Ähnlichkeiten zwischen vier Punkten



Bezeichnung der Punkte (Rangfolgen):

1: <abdecf>	11: <abcdef>	21: <abefdc>
2: <abedcf>	12: <afbcde>	22: <abedfc>
3: <abdfce>	13: <afbecd>	23: <abecfd>
4: <abfdce>	14: <afbedc>	24: <abfcde>
5: <afbdce>	15: <abdcfe>	25: <abfecd>
6: <afbdec>	16: <abdefc>	26: <abfedc>
7: <abdfec>	17: <abecdf>	27: <abdcfe>
8: <abdcef>	18: <afbcde>	28: <abcfde>
9: <abfdec>	19: <abfcde>	29: <abcfed>
10: <abcedf>	20: <abefcd>	30: <abcefd>

Die Rangordnung ist ansteigend.

Unterstrichene Rangfolgen sind im R^1 darstellbar.

Verteilung der Abstände	i									
	0	1	2	3	4	5	6	7	Σ	
1. zwischen einer Rangfolge und allen anderen										
a. für Rangfolgen 4, 9, 19, 26, 25, 24	1	4	7	8	6	3	1	0	30	
b. für Rangfolgen 1, 8, 6, 18, 5, 11, 2, 17, 14, 12, 13, 10	1	3	5	6	6	5	3	1	30	
c. für Rangfolgen 16, 27, 15, 22, 30, 23	1	3	5	7	7	5	2	0	30	
d. für Rangfolgen 3, 28, 7, 20, 29, 21	1	3	6	8	7	4	1	0	30	
2. in der Gesamtheit	30	48	84	105	96	66	30	6	465	
3. zu den darstellbaren "D"	5	13	27	33	29	20	10	3	140	
4. zu den nicht darstellbaren	25	45	83	104	94	64	29	6	450	
5. zwischen den darstellbaren	5	3	1	1	2	2	1	0	15	
6. Minimalabstände zu den darstellbaren	5	10	10	5	0	0	0	0	30	
7. Minimalabstände zur Menge {3, 4, 5}	3	6	8	7	4	2	0	0	30	

Tabelle 2 : Verteilungen $P_{r(\theta)}(\{r(x)\}_i)$ nach (4.1) ; $\{r(x)\}_i$ wie unten

$r(\theta)$	i :							
	0	1	2	3	4	5	6	7
4, 9, 19, 26, 25, 24	$\frac{7}{123}$	$\frac{24}{123}$	$\frac{35}{123}$	$\frac{32}{123}$	$\frac{18}{123}$	$\frac{6}{123}$	$\frac{1}{123}$	0
1, 8, 6, 18, 5, 11, 2, 17, 14, 12, 13, 10	$\frac{8}{135}$	$\frac{21}{135}$	$\frac{30}{135}$	$\frac{30}{135}$	$\frac{24}{135}$	$\frac{15}{135}$	$\frac{6}{135}$	$\frac{1}{135}$
16, 27, 15, 22, 30, 23	$\frac{7}{111}$	$\frac{18}{111}$	$\frac{25}{111}$	$\frac{28}{111}$	$\frac{21}{111}$	$\frac{10}{111}$	$\frac{2}{111}$	0
3, 28, 7, 20, 29, 21	$\frac{7}{117}$	$\frac{18}{117}$	$\frac{30}{117}$	$\frac{32}{117}$	$\frac{21}{117}$	$\frac{8}{117}$	$\frac{1}{117}$	0

θ : Die "wahre", zugrundeliegende Rangfolge
 $\{r(x)\}_i = \{r(x) \mid S(r(\theta), r(x)) = i\}$

a, b, c, d, e und f seien dabei die Ähnlichkeiten zwischen den Objekten A, B, C, D :

	A	B	C	D
A	0	a	b	c
B	a	0	d	e
C	b	d	0	f
D	c	e	f	0

wobei $a = S(A, B)$ etc.

Im obigen Graphen sind diejenigen Ecken hervorgehoben, in denen die im R^1 darstellbaren Rangfolgen stehen.

4.4 Verteilungen

Im obigen Graphen können leicht die Verteilungen der Abstände (in Permutationsschritten) zwischen den wesentlichen Rangfolgen durch Auszählen bestimmt werden.

Tabelle 1 enthält einige dieser Verteilungen. i bezeichnet die Größe der Abstände zwischen zwei Rangfolgen in Permutationsschritten. In den Feldern der Tabelle stehen die Häufigkeiten $h(i)$, mit denen die Abstände auftreten.

Die Abstände zu einer Menge von Rangfolgen sind alle Abstände zwischen je einem Element der Menge und allen Rangfolgen der Gesamtmenge.

Minimalabstände zu einer Menge sind die Abstände aller Rangfolgen zum jeweils nächstliegenden Element der Menge (vgl. Tabelle 1).

Auf Grundlage dieser Verteilungen können nun verschie-

dene Hypothesen getestet werden:

1. Angenommen, ein Proband absolviert ein psychologisches Testverfahren, dessen Ergebnis eine Rangskala, x , mit sechs Elementen ist. X sei eine Zufallsvariable über der Menge aller möglichen Rangfolgen aus sechs Elementen, θ sei die zugrundeliegende "wahre" Rangfolge. Ersetzt man x durch den Repräsentanten $r(x)$ der Klasse der Rangfolgen, die mit x bis auf Namensvertauschung der Punkte identisch sind, so gibt der obige Graph den Wertebereich von $r(x)$ wieder. $r(x)$ sei $N(r(\theta), 1)$ -verteilt.

Die Hypothese H_0 : " $r(x)$ ist Element der Menge der darstellbaren Rangfolgen" wird dann mit Fehlerwahrscheinlichkeit $\alpha = 5\%$ abgelehnt, falls der Permutationsabstand S zwischen $r(x)$ und dem nächstliegenden Element der Menge der darstellbaren Rangfolgen 2 oder mehr beträgt. Der Fehler zweiter Art läßt sich leicht für jedes nicht darstellbare $r(\theta)$ aus der $N(r(\theta), 1)$ -Verteilung bestimmen.

2. Für jede der Rangfolgen $r(x)$ aus Beispiel 1. läßt sich ein Konfidenzbereich zum Niveau $1 - \alpha = 95\%$ angeben, in dem die "wahre" Rangfolge mit Wahrscheinlichkeit von 95% liegen muß. Dieser Bereich besteht nach der unter 1. angenommenen Verteilung aus der Rangfolge $r(x)$ und allen Rangfolgen, die nicht weiter als einen Permutationsschritt von $r(x)$ entfernt liegen.
3. Verändert man die H_0 -Hypothese und / oder die Verteilungsannahme, so ergeben sich verschiedene andere Hypothesentests:

Sei die Hypothese H_0 : "Die wahre Rangfolge ist dar-

stellbar und enthält die Ungleichung $c < e$." Diese Eigenschaften besitzen im obigen Graphen die Rangfolgen 3, 4 und 5. Unterstellt man, die $r(x)$ seien uniformverteilt, so könnte die H_0 -Hypothese mit $\alpha = 5\%$ für kein x abgelehnt werden, da die von $\{3, 4, 5\}$ entferntesten Rangfolgen 23 und 17 zusammen die Wahrscheinlichkeit $2/30 \approx 0,066$ haben.

Realistischer als die Verteilungsannahme, daß die $r(x)$ uniformverteilt sind, ist die Annahme der Dreiecksverteilung. Das heißt, für die "wahre" Rangfolge $r(\theta)$ nimmt die Wahrscheinlichkeit $P(r(x))$ mit wachsendem Permutationsabstand $S(r(\theta), r(x))$ linear ab.

Konstruktion: Die Häufigkeiten $h_{r(\theta)}(i)$ aus Tabelle 1 werden multipliziert mit

$$(i_{\max} + 1 - i); \text{ mit } i_{\max} = \max_i$$

und durch die Gesamtsumme dieser Produkte dividiert:

$$P_{r(\theta)}(\{r(x)\}_i) = \frac{h_{r(\theta)}(i) (i_{\max} + 1 - i)}{\sum_i h_{r(\theta)}(i) (i_{\max} + 1 - i)} \quad (4.1)$$

$$\text{mit } \{r(x)\}_i = \{r(x) \mid \min_{r(\theta) \in \{3, 4, 5\}} S(r(\theta), r(x)) = i\}$$

d.h. diese Mengen bestehen aus den $r(x)$, die mit Minimalabstand i von $r(\theta)$ entfernt liegen.

Motivation für die Hypothese $H_0: r(\theta) \in \{3, 4, 5\}$:

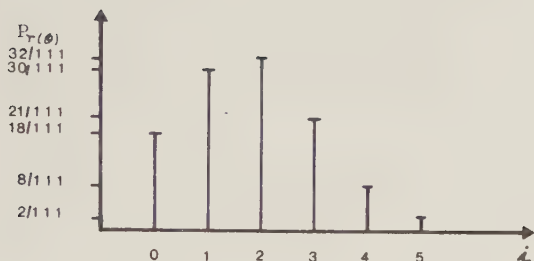
Nach Tabelle 1, Verteilung 7. gibt es sechs ver-

schiedene Minimalabstände i zur Menge $\{3,4,5\}$.
 Geht man nun davon aus, daß die Wahrscheinlichkeiten für zwei Rangfolgen mit gleichem Minimalabstand zur Menge $\{3,4,5\}$ gleich groß sind und davon, daß die Differenzen der Wahrscheinlichkeiten zweier Mengen mit Abstand i und $i + 1$ von $\{3,4,5\}$ für alle i gleich bleiben, so kann man für Mengen $\{r(x)\}_i$ von Elementen mit gleichem Minimalabstand zu $\{3,4,5\}$ folgende Verteilung konstruieren:

$$\begin{array}{cccccc}
 i = & 0 & 1 & 2 & 3 & 4 & 5 \\
 \hline
 p_i = & \frac{6}{21} & \frac{5}{21} & \frac{4}{21} & \frac{3}{21} & \frac{2}{21} & \frac{1}{21}
 \end{array}$$

Die Hypothese H_0 wird hier für alle $r(x)$ mit $S_{\min} = 5$ zum Niveau $\alpha = 5\%$ zurückgewiesen.

Diese Verteilung berücksichtigt nicht die Größen der Mengen $\text{card}\{r(x)\}_i$, d.h. die Anzahl der Rangfolgen, die in Mengen mit gleichem Minimalabstand von $r(\theta)$ liegen. Die Größe einer Menge wird aber einen Einfluß auf die Wahrscheinlichkeit haben, ein Element aus ihr zu ziehen. Wir wollen deshalb die Wahrscheinlichkeiten p_i mit der Anzahl der Elemente der Mengen $\{r(x)\}_i$ gewichten, vgl. Tabelle 1, Zeile sieben. Mittels Division durch ihre Gesamtsumme normieren wir diese Terme und erhalten eine Verteilung, deren Dichte die Form hat:



$$P_r(\theta) (\{r(x)\}_i) = \frac{p_i \cdot \text{card}(\{r(x)\}_i)}{\sum_i p_i \cdot \text{card}(\{r(x)\}_i)}$$

(Vgl. auch Tabelle 2, die die entsprechenden Verteilungen um einzelne $r(\theta)$ wiedergibt.)

Unter dieser Verteilungsannahme läßt sich wiederum leicht für jedes $r(\theta) \notin \{3, 4, 5\}$ der Fehler zweiter Art berechnen.

4. In diesem Beispiel sind fünf Klassen im R^1 darstellbar, die restlichen im R^2 . Bei größerer Anzahl von Punkten ist ein Test über die Darstellbarkeit in einem Raum gegebener Dimension gegen eine andere Dimension möglich.
5. Für Rangfolgen mit fehlenden Werten können auf einfache Weise Permutations-Abstände zu anderen Rangfolgen bestimmt werden. Es wird die Anzahl der Schritte ermittelt, die notwendig ist, um die gemessenen Unähnlichkeiten in die entsprechenden der

zu vergleichenden Rangfolge zu überführen. Der Abstand z.B. zwischen den Rangfolgen $\langle a, b, c, d, e, f \rangle$ und $\{ \langle b, d, c, e \rangle, \langle f, a \rangle \}$ ist dann 6, da ein Schritt notwendig ist, a und c sowie fünf Schritte nötig sind, um a und f zu vertauschen. Die obige Schreibweise soll aussagen, daß in der zweiten Rangfolge die Vergleiche zwischen $\langle b, d, c, e \rangle$ und f sowie $\langle b, d, c, e \rangle$ und a fehlen. Es ist dadurch also möglich, auch missing-data-Hypothesentests und -Konfidenzbereichsschätzungen durchzuführen.

Die Beispiele zeigen, daß es nach Berechnung der Verteilungen der Permutationsabstände möglich ist, Hypothesentests über die Entfernung einer gemessenen Rangfolge zu darstellbaren zu berechnen und auch deren Schärfe zu bestimmen.

Beim Prüfen von Modellen ist die Beschränkung des Fehlers erster Art nicht entscheidend, wie oben erörtert wurde. Vielmehr müssen die Fehler erster und zweiter Art beide gleichermaßen berücksichtigt werden. Das Risi kokonzept der Statistik bietet die Möglichkeit, Meßwerte für die Gesamtheit der zu erwartenden Fehler zu berechnen. Dazu wird eine Verlustfunktion $L(\varphi(x), \theta)$ benötigt. $\varphi(x)$ ist die Entscheidungsfunktion, die jeder gemessenen Rangfolge x eine geschätzte, z.B. die der NMDS-Konfiguration zuweist. L gibt bei vorliegendem Meßwert x und bekanntem θ den Verlust an, der entsteht, wenn $\varphi(x)$ von θ abweicht. Multipliziert man $L(\varphi(x), \theta)$ mit $P_{\theta}(x)$, also der Wahrscheinlichkeit des Auftretens von x falls θ die wahre Rangfolge ist, so erhält man nach Summation über x die Risikofunktion.

$$R(\varphi, \theta) = \sum_x L(\varphi(x), \theta) \cdot P_{\theta}(x) \quad (4.2)$$

Dabei wird von nicht-randomisierten Entscheidungsfunktionen ausgegangen, da Randomisierung beim Schätzen von NMDS-Modellen nicht üblich und wenig sinnvoll ist.

Beispiel:

6. Die Entscheidungsfunktion sei hier definiert als $\varphi(r(x)) = r(x)$, d.h. man entscheidet sich für die gerade gemessene Rangfolge und nicht etwa für eine darstellbare, es sei denn, die gerade gemessene ist selbst darstellbar. Eine Verlustfunktion sei das Permutationsschrittmaß $S(\varphi(x), \theta)$, d.h. die minimale Anzahl der Permutationsschritte, die zwischen $\varphi(x)$ und θ liegen. Dann ist

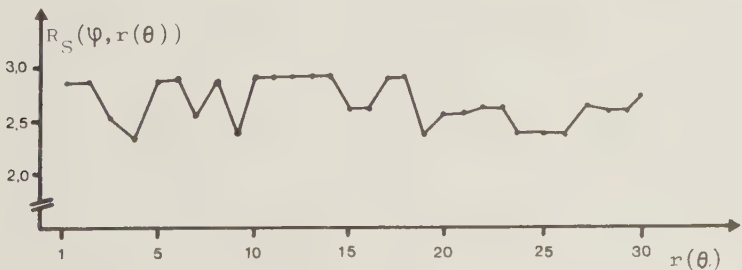
$$R_S(\varphi, \theta) = \sum_x S(\varphi(x), \theta) \cdot p_\theta(x).$$

Wir wollen uns wieder an die Verteilung (4.1) und Bezeichnungsweise von Beispiel 3. halten und setzen $r(x)$ und $r(\theta)$ anstelle von x und θ .

Aus Tabelle 2 berechnet man z.B.

$$R_S(\varphi, 4) = \sum_{i=1}^6 (7 - i) \cdot p_i = 2,42276.$$

Trägt man $R_S(\varphi, r(\theta))$ über $r(\theta)$ ab, so erhält man



$$R_S(\psi, 1) = 2,85185$$

$$R_S(\psi, 15) = 2,68468$$

$$R_S(\psi, 3) = 2,59829$$

Wählt man anstelle der Verlustfunktion S das normierte Permutationsmaß s (vgl. 3.8), so ergibt sich

$$R_s(\psi, 4) = 0,40379$$

$$R_s(\psi, 1) = 0,40740$$

$$R_s(\psi, 15) = 0,44744$$

$$R_s(\psi, 3) = 0,43304$$

Je größer die Varianz von $r(X)$, desto größer wird bei fester Verlust- und Entscheidungsfunktion das Risiko. Wir werden im nächsten Kapitel einige alternative Verteilungen für X definieren.

In einem Zufallsexperiment, in dem eine Verteilung H der θ bekannt ist, wie es im nächsten Kapitel beschrieben wird, kann das gemittelte oder "Bayes"-Risiko bestimmt werden:

$$R(\psi, H) = \sum_{\theta} R(\psi, \theta) \cdot h(\theta)$$

$h(\theta)$ ist die Dichtefunktion von θ .

Ist θ uniformverteilt, so wird in Beispiel 6.

$$R_S(\psi, H) = \frac{6}{30} R(\psi, 4) + \frac{12}{30} R(\psi, 1) + \frac{6}{30} R(\psi, 15) +$$

$$\frac{6}{30} R(\psi, 3) = 2,681922$$

und

$$R_s(\varphi, H) = 0,41981$$

Das heißt, gemittelt über alle x und θ beträgt der Verlust etwa 2,68 Permutationsschritte oder ca. 41% vom maximalen Permutationsabstand.

Beispiel:

7. Die H_0 -Hypothese heiße hier: $r(\theta) \in \{1, 2, 3, 4, 5\} = D$
 Gegenhypothese: $r(\theta) = r_1(\theta) \notin \{1, 2, 3, 4, 5\}$ d.h. $r(\theta)$ nimmt einen bestimmten Wert außerhalb der Menge der darstellbaren Rangfolgen an.

Sei $\varphi(r(x))$ definiert als die Entscheidung für die nächstliegende im R^1 darstellbare Rangfolge, $d(x)$:

$$\varphi(r(x)) = d(x)$$

$$\text{mit } S(d(x), r(x)) = \min_{d(x) \notin D} S(d(x)', r(x)),$$

$$d(x) \in D$$

Eine Verlustfunktion sei definiert als

$$L_{\min}(d(x), r(x)) = s(d(x), r(x))$$

das ist die normierte minimale Anzahl von Permutationsschritten zwischen θ und dem nächstliegenden Element der Menge $\{1, 2, 3, 4, 5\}$ (siehe Tabelle 1, Zeile 7).

Man errechnet daraus:

$r(\theta) \in$	$R_S(\psi, \theta)$	$R_s(\psi, \theta)$
$\{1, 2, 3, 4, 5\}$	0	0
$\{6, 7, 8, 9, 15, 16, 17, 18, 19, 22\}$	1	0,33333
$\{10, 11, 12, 14, 21, 23, 24, 26, 27, 28\}$	2	0,66666
$\{13, 20, 25, 29, 30\}$	3	1

$$R_S(\psi, H) = 1,5$$

$$R_s(\psi, H) = 0,49999 \quad , \quad \theta \text{ uniformverteilt}$$

In diesem Beispiel ist der Verlust, der bei einer falschen Entscheidung entsteht, unabhängig von der konkret gewählten Rangfolge aus der Menge der darstellbaren; es geht nur die Entfernung von $r(\theta)$ zu dieser Menge in die Verlustfunktion ein. Dies ist für herkömmliche NMDS-Modellschätzungen aber realistisch, da im allgemeinen keine Hypothese über eine bestimmte Rangfolge gemacht werden kann und nur der Unterschied zwischen den Daten und der ausgesuchten "Minimalen-Streß-Konfiguration" ins Gütekriterium, d.h. in die Streßfunktion, eingeht. Weiter ist zu bemerken, daß die Entscheidungsfunktion $\psi(r(x))$ immer die nächstliegende darstellbare Rangfolge auswählt, ohne jemals eine andere, nicht darstellbare, zu berücksichtigen. Auch das entspricht weitgehend dem herkömmlichen Vorgehen insbesondere bei geringen Zahlen (n) von Objekten, wo ein geringer Streß für alle möglichen gemessenen Rangfolgen eine NMDS-Darstellung "zuläßt".

Die Bayes-Risiken in den Beispielen 6. und 7. erläutern keinen Vergleich der Entscheidungsfunktionen, da die Hypothesen und Verlustfunktionen verschieden sind.

Im nächsten Kapitel sollen Entscheidungsfunktionen

über das Bayes-Risiko miteinander verglichen werden.

Der Einfachheit und Darstellbarkeit halber wurden hier nur die wesentlichen Rangfolgen betrachtet, die von der inhaltlichen Bestimmtheit der Punkte absehen und nur die geometrische Form einer räumlichen Darstellung der repräsentierten Rangfolgen bewahren. Diese Einschränkung wird im folgenden aufgehoben.

Vorher sei noch erwähnt, daß die Elemente der Klassen der wesentlichen Rangfolgen für $n > 3$ nicht dicht liegen. Für $n = 4$ ist der minimale Abstand zwischen zwei verschiedenen Rangfolgen einer Klasse gleich zwei, da sich durch Vertauschung zweier Punktnamen vier Rangplätze verändern.

Die weitere Aufgabe besteht nun darin, Aussagen über die Permutationsdistanzen für höhere Punktezahlen und für Rangfolgen, die in Räumen höherer Dimension darstellbar sind, zu machen. Dadurch können die Ergebnisse dieses Abschnitts verallgemeinert werden. Zunächst soll das Testproblem statistisch formuliert werden.

4.5 Statistische Formulierung des Testproblems

Da für endliche Punktemengen auch die Anzahl der Rangfolgen endlich ist, kann hier auf Grenzwerteigenschaften völlig verzichtet werden. Die zugrundeliegende Menge von Rangfolgen $(\mathcal{X})$ bildet bezüglich der Mengenverknüpfungen eine Algebra $(\mathcal{B})$. Darauf sei ein Wahrscheinlichkeitsmaß (P_θ) definiert. Der Parameter kann selbst wieder eine Zufallsvariable sein aus einem Wahrscheinlichkeitsraum (Θ) mit Algebra $(\mathcal{G})$ und W.-Maß $(\mathcal{X})$.

Sei $X \in (\mathcal{X}, \mathcal{B}, P_\theta)$ eine gemessene Rangfolge. $\varphi(x) \in (\Delta, \mathcal{G})$ (Algebra, Maß), sei eine Entscheidung für eine Rangfolge, z.B. eine darstellbare Rangfolge. θ kann z.B. die "wahre" zugrundeliegende Rangfolge sein. Der Fehler, der durch die Entscheidung $\varphi(x)$ gemacht wurde, wird dann durch die Verlustfunktion $L(\varphi(x), \theta)$, z.B. das Permutationsmaß $S(\varphi(x), \theta)$, bezeichnet. Mittelt man die Verlustfunktion über alle möglichen Stichprobenergebnisse x und alle möglichen Entscheidungen $\varphi(x) = a \in \Delta$, so erhält man die "Risikofunktion"

$$R(\varphi, \theta) = \sum_x \sum_{\Delta} L(a, \theta) \cdot P_\theta(x) \cdot P(a).$$

Beim Testproblem

$H_0: \theta \in \Theta_0$ (die "wahre" Rangfolge ist aus Θ_0 , z.B. der Menge der im $\mathbb{R}^p$ darstellbaren)

$H_1: \theta \in \Theta_1$ (die "wahre" Rangfolge ist aus Θ_1 , z.B. der Menge der nicht darstellbaren)

ist für nicht randomisierte φ eine mögliche Risikofunktion gegeben durch

$$R(\varphi, \theta) = \sum_x (S(a_1, \theta) \cdot I_{\Theta_0}(\theta) \cdot P_\theta(x) + S(a_0, \theta) \cdot I_{\Theta_1}(\theta) \cdot P_\theta(x)) \quad (4.3)$$

$$\cdot I_{\Theta_1}(\theta) \cdot P_\theta(x))$$

$$\text{mit } I_{\Theta_0}(\theta) = \begin{cases} 1 & \text{falls } \theta \in \Theta_0 \\ 0 & \text{sonst} \end{cases}$$

a_1 = Entscheidung für H_1

a_0 = Entscheidung für H_0

Der erste Summand in der Klammer entspricht dem Ver-

lust beim Fehler erster Art multipliziert mit seiner Wahrscheinlichkeit, der zweite Summand dem Verlust beim Fehler zweiter Art mal Wahrscheinlichkeit. a_0 wird i.a. die Rangfolge mit bestem Streß im betrachteten R^P bedeuten. $R(\varphi, \theta)$ kann für jedes θ andere Werte annehmen. Hat man eine Annahme über die Verteilung H der θ , so kann man das Risiko über alle θ mitteln (Bayes-Risiko):

$$R(\varphi, H) = \sum_{\theta} R(\varphi, \theta) \cdot h(\theta) \quad (4.4)$$

Zur Lösung dieser Probleme werden daher benötigt:

1. Die Verteilung der Permutationsabstände zwischen einer Rangfolge und allen anderen. Diese Verteilung braucht man für die Entscheidung über die Hypothese $H_0: \theta = \theta_0$, d.h. der Hypothese, daß nur eine mögliche "wahre" Rangfolge existiert.
2. Die Verteilung der Permutationsabstände zwischen einer Menge von Rangfolgen Θ_0 und allen anderen. Diese Verteilung braucht man für die Entscheidung über die Hypothese $H_0: \theta \in \Theta_0$. Für die Gegenhypothese $H_1: \theta \in \Theta_1$ braucht man ferner die Verteilung der Permutationsabstände zwischen Θ_1 und allen anderen. Θ_0 wird im folgenden als Menge der darstellbaren Rangfolgen spezifiziert. Diese Festlegung ist aber allgemein nicht notwendig, vielmehr können bei entsprechender Definition von Θ_0 Hypothesen über andere Eigenschaften der Rangfolgen, z.B. Fixierungen von Teilfolgen und andere "Constraints" formuliert werden.

5. Kapitel

Verteilungen von Rangfolgen, Risikofunktionen

Wegen der endlichen Zahl von Rangfolgen für endliche Punktezahlen und der relativ einfachen Möglichkeit, die Permutationsdistanz zwischen zwei Rangfolgen zu berechnen, erscheint die Aufgabe, die Verteilungen der Distanzen von Rangfolgen mit spezieller Eigenschaft untereinander und zu anderen Mengen von Rangfolgen zu bestimmen, ein lösbares Problem der Kombinatorik, oder doch in Form einer geschlossenen Formel angebar zu sein. Diese Herangehensweise stößt jedoch schnell auf so erhebliche Schwierigkeiten, daß die Durchführung für größere Punktezahlen und Dimensionen technisch nicht möglich ist. Es wird daher auf die exakte Darstellung verzichtet und ein Simulationsprogramm entwickelt, das die benötigten Verteilungen mit hinreichender Genauigkeit schätzt.

5.1 Verteilung der Distanzen zwischen darstellbaren Rangfolgen

Nach dem Zufallsprinzip werden Paare von Konfigurationen mit fester Punkte- und Dimensionszahl erzeugt sowie der Permutationsabstand jedes der Paare berechnet. Der Umfang einer solchen Stichprobe aus der Menge von Paaren darstellbarer Rangfolgen ist 1.000.

Für jeden Permutationsabstand werden unabhängig voneinander zwei neue Zufallskonfigurationen generiert. Alternativ wäre vorstellbar gewesen, eine gewisse Anzahl k (z.B. 100) von Zufallskonfigurationen zu erzeugen und die Menge der paarweisen Abstände, also $k \cdot (k - 1) / 2$, zu berechnen. Ein solches Vorgehen

würde jedoch nicht die Unabhängigkeit der einzelnen Ziehungen voneinander garantieren: Berechnet man die Distanzen von jeder der k Rangfolgen zu nur drei vorher fixierten, so sind aus diesen $k - 3$ Distanzen sämtliche $k \cdot (k - 1) / 2 = 3k$ restlichen eindeutig bestimmbar. Eine solche - abhängige - Stichprobe ist zur Bestimmung der Verteilungen nicht geeignet.

Jede Zufallskonfiguration entstand durch Auswahl von n Punkten über die Uniformverteilung mit Wertebereich $[0, 1]^p$ (p -dimensionales Intervall $[0, 1]$).

D.h. jede der p Komponenten jedes Punktes wurde nach dem Monte-Carlo-Verfahren aus einer $[0 - 1]$ -Uniformverteilung gezogen. Dazu wurde die FORTRAN-Routine RANF benutzt, die nach dem Prinzip der multiplikativen Kongruenz arbeitet. (vgl. Programm 1 im Anhang)

Jede im R^p darstellbare Rangfolge kann in potentiell unendlich viele räumliche R^p -Konfigurationen abgebildet werden. Die Menge der möglichen Abbildungen jeder darstellbaren Rangfolge bildet einen konvexen Kegel (vgl. Coombs (5)). Nun sind diese Kegel für verschiedene Rangfolgen auch bei gleicher Punkte- und Dimensionszahl verschieden groß. Daher bedeutet die Einhaltung der Uniformverteilung bei der Konstruktion der Punkte der Zufallskonfigurationen nicht, daß damit auch über der Menge der darstellbaren Permutationen eine Uniformverteilung zugrundegelegt wird. Da wir die Verteilung der darstellbaren Rangfolgen jedoch unter der Hypothese der Darstellbarkeit benutzen, d.h. annehmen werden, daß eine gegebene Rangfolge eine Rangfolge von Distanzen einer räumlichen Konfiguration ist, ist hier die Uniformverteilung über der Menge aller räumlichen Konfigurationen und nicht über der der darstellbaren Rangfolgen gerade gewünscht. Mit anderen Worten: Den Daten wird in der H_0 -Hypothe-

se gerade unterstellt, sie seien zufällige Abweichungen von darstellbaren Rangfolgen bzw. von Distanzfolgen räumlicher Konfigurationen. Dann wird ihre Verteilung aber gerade durch solche Rangfolgen geschätzt, die nach der Größe ihres konvexen Kegels gewichtet sind, um die Häufigkeit ihres Auftretens als Konfigurationen zu berücksichtigen.

Für die Berechnung der Permutationsdistanzen wurde eine Routine geschrieben, die nach dem Prinzip der Umordnung beider Rangfolgen arbeitet. Beide Rangfolgen werden simultan so umgeordnet, daß eine der beiden zur natürlichen Rangfolge wird. Bei der anderen wird für jedes Glied dann die Anzahl der Folgeglieder festgestellt, die bezüglich der natürlichen Reihenfolge kleiner sind, in der umgeformten Rangfolge aber an höherer Stelle stehen. Die Summe dieser Zahlen ist der Permutationsabstand S .

Beispiel:

Folgen $F_1 = \langle a, d, g, h, c, b, f, e \rangle$

und $F_2 = \langle b, f, c, d, a, e, g, h \rangle$

Simultanes Umordnen, so daß F_1 in natürlicher Reihenfolge steht:

$F_1' = \langle a, b, c, d, e, f, g, h \rangle$

$F_2' = \langle b, e, a, f, h, g, c, d \rangle$

Überprüfung der Glieder von Folge F_2' :

<u>Verletzung der natürli-</u> <u>chen Reihenfolge durch:</u>	<u>Anzahl der Ver-</u> <u>letzungen</u>
1. b.....a.....	1
2. e.....a,c,d.....	3
3. a.....-.....	0
4. f.....c,d.....	2
5. h.....g,c,d.....	3
6. g.....c,d.....	2
7. c,d.....-.....	0

$$\Sigma = 11 = S(F_1, F_2)$$

Dieses Verfahren wurde gewählt, weil das zwar schnellere aber speicheraufwendigere Matrix-Verfahren (vgl. Kendall, (18)) für größere Punktezahlen ($n > 15$) technisch nicht durchführbar ist (vgl. (17)).

Die Verteilungen für verschiedene Punkte- und Dimensionenzahlen wurden normiert zum Wertebereich $[0, +1]$, d.h. es wird der normierte Permutationsabstand s betrachtet, das direkt mit Kruskal's Streß und Guttman's Alienationskoeffizienten verglichen werden kann.

Tabelle 5.1 zeigt Werte der empirischen Verteilungen (absolute kumulierte Häufigkeiten) der normierten Permutationsabstände zwischen den Elementen aus der Menge der darstellbaren Rangfolgen. Die Zahl der Punkte variiert zwischen 4 und 40 für 1 bis 5 Dimensionen. Für 4 - 15 Punkte wurden 1000 Abstände gemessen, für 20 - 35 Punkte aus Kostengründen 100, für 40 Punkte 30 Abstände.

Tabelle 5.1: Empirischen Häufigkeiten von Abständen zwischen Elementen aus D

[illegible]

NORM.PERMUTATIONSABSTAND

PKT	DIM	.15	.20	.25	.30	.35	.38	.41	.44	.47	.50	.53	.56	.59	.62	.65	.70	.75	.80	.85	1.00
9	1	0	0	0	0	2	22	86	206	381	557	701	809	889	938	968	990	997	999	1000	1.000
9	2	0	0	0	0	5	36	104	215	375	533	664	792	863	934	969	991	999	1000	1000	1000
9	3	0	0	0	2	12	32	86	192	342	493	648	792	895	953	983	997	999	1000	1000	1000
9	4	0	0	0	2	8	37	89	192	326	495	659	801	905	962	984	996	1000	1000	1000	1000
9	5	0	0	0	0	7	30	76	178	335	500	674	829	923	974	988	999	1000	1000	1000	1000
10	1	0	0	0	0	0	14	55	171	344	545	715	863	919	966	986	996	999	1000	1000	1000
10	2	0	0	0	0	3	23	67	170	331	534	725	831	904	952	974	996	1000	1000	1000	1000
10	3	0	0	0	0	9	27	85	196	348	536	695	855	932	969	987	999	1000	1000	1000	1000
10	4	0	0	0	0	5	26	74	156	330	529	714	849	935	977	996	1000	1000	1000	1000	1000
10	5	0	0	0	0	3	24	71	176	332	515	710	850	948	984	997	1000	1000	1000	1000	1000
11	1	0	0	0	0	1	5	36	145	324	543	749	869	940	974	993	999	1000	1000	1000	1000
11	2	0	0	0	0	1	13	59	160	322	514	703	839	924	958	986	997	999	1000	1000	1000
11	3	0	0	0	0	3	12	58	166	324	516	709	855	939	976	996	1000	1000	1000	1000	1000
11	4	0	0	0	0	0	7	47	129	285	486	689	841	935	978	995	1000	1000	1000	1000	1000
11	5	0	0	0	0	1	12	49	122	294	498	723	887	959	984	997	1000	1000	1000	1000	1000
12	1	0	0	0	0	0	3	26	126	316	561	723	856	929	969	987	997	1000	1000	1000	1000
12	2	0	0	0	0	4	18	57	148	306	520	710	859	932	968	988	999	1000	1000	1000	1000
12	3	0	0	0	0	3	13	42	139	304	511	694	863	940	985	996	1000	1000	1000	1000	1000
12	4	0	0	0	0	2	11	46	134	301	492	693	870	950	984	998	1000	1000	1000	1000	1000
12	5	0	0	0	0	0	3	32	117	290	499	731	903	970	994	1000	1000	1000	1000	1000	1000
13	1	0	0	0	0	0	1	20	121	318	555	764	895	955	987	995	1000	1000	1000	1000	1000
13	2	0	0	0	0	1	6	29	121	291	518	707	853	933	971	988	999	1000	1000	1000	1000
13	3	0	0	0	0	2	3	35	117	289	541	723	877	946	987	1000	1000	1000	1000	1000	1000
13	4	0	0	0	0	1	7	36	105	270	506	738	888	962	989	999	1000	1000	1000	1000	1000
13	5	0	0	0	0	0	3	22	96	247	488	726	893	967	996	1000	1000	1000	1000	1000	1000

5.1 Fortsetzung

PPT DIM	NORM. PERMUTATIONSABSTAND																			
	.15	.20	.25	.30	.35	.38	.41	.44	.47	.50	.53	.56	.59	.62	.65	.70	.75	.80	.85	1.00
14 1	0	0	0	0	0	1	10	77	260	521	741	888	969	990	998	1000	1000	1000	1000	1000
14 2	0	0	0	0	0	6	32	111	314	530	729	878	950	990	996	1000	1000	1000	1000	1000
14 3	0	0	0	0	0	0	30	108	261	476	713	883	959	989	998	1000	1000	1000	1000	1000
14 4	0	0	0	0	0	1	25	98	273	531	744	888	972	993	998	1000	1000	1000	1000	1000
14 5	0	0	0	0	1	3	21	90	246	527	758	921	985	998	1000	1000	1000	1000	1000	1000
15 1	0	0	0	0	0	0	11	72	259	536	770	901	965	990	998	1000	1000	1000	1000	1000
15 2	0	0	0	0	1	2	20	103	280	521	744	878	953	982	999	1000	1000	1000	1000	1000
15 3	0	0	0	0	1	3	27	105	291	519	752	903	979	991	998	1000	1000	1000	1000	1000
15 4	0	0	0	0	0	1	26	89	232	512	761	913	972	996	999	1000	1000	1000	1000	1000
15 5	0	0	0	0	0	4	22	87	230	509	769	925	984	999	1000	1000	1000	1000	1000	1000
20 1	0	0	0	0	0	0	0	2	16	49	91	98	100	100	100	100	100	100	100	100
20 2	0	0	0	0	0	0	0	2	22	52	76	94	97	100	100	100	100	100	100	100
20 3	0	0	0	0	0	0	0	3	23	55	82	95	100	100	100	100	100	100	100	100
20 4	0	0	0	0	0	0	0	5	21	51	83	94	100	100	100	100	100	100	100	100
20 5	0	0	0	0	0	0	1	1	14	42	82	99	100	100	100	100	100	100	100	100
25 1	0	0	0	0	0	0	0	2	13	43	76	94	98	100	100	100	100	100	100	100
25 2	0	0	0	0	0	0	0	1	15	58	83	97	98	100	100	100	100	100	100	100
25 3	0	0	0	0	0	0	0	5	17	56	88	97	100	100	100	100	100	100	100	100
25 4	0	0	0	0	0	0	0	2	16	50	85	93	99	100	100	100	100	100	100	100
25 5	0	0	0	0	0	0	0	1	13	55	90	99	100	100	100	100	100	100	100	100
30 1	0	0	0	0	0	0	0	0	8	59	90	98	99	100	100	100	100	100	100	100
30 2	0	0	0	0	0	0	0	1	15	49	85	96	100	100	100	100	100	100	100	100
30 3	0	0	0	0	0	0	0	0	11	41	73	94	100	100	100	100	100	100	100	100
30 4	0	0	0	0	0	0	0	0	11	51	85	97	100	100	100	100	100	100	100	100
30 5	0	0	0	0	0	0	0	0	10	44	86	99	100	100	100	100	100	100	100	100

5.1 Fortsetzung

		NORM. PERMUTATIONSABSTAND																			
PKT DIM		.15	.20	.25	.30	.35	.38	.41	.44	.47	.50	.53	.56	.59	.62	.65	.70	.75	.80	.85	1.00
35	1	0	0	0	0	0	0	0	2	7	56	93	100	100	100	100	100	100	100	100	100
35	2	0	0	0	0	0	0	0	0	6	55	88	99	100	100	100	100	100	100	100	100
35	3	0	0	0	0	0	0	0	1	10	51	85	99	100	100	100	100	100	100	100	100
35	4	0	0	0	0	0	0	0	1	4	43	87	100	100	100	100	100	100	100	100	100
35	5	0	0	0	0	0	0	0	0	9	47	87	99	100	100	100	100	100	100	100	100
40	1	0	0	0	0	0	0	0	0	4	20	28	30	30	30	30	30	30	30	30	30
40	2	0	0	0	0	0	0	0	0	7	19	29	30	30	30	30	30	30	30	30	30
40	3	0	0	0	0	0	0	0	0	7	17	25	30	30	30	30	30	30	30	30	30
40	4	0	0	0	0	0	0	0	0	1	18	27	30	30	30	30	30	30	30	30	30
40	5	0	0	0	0	0	0	0	0	3	16	28	30	30	30	30	30	30	30	30	30

Die empirischen Verteilungen approximieren die Normalverteilungen mit $\mu = 0.5$ und $\sigma^2 = \frac{2m+5}{18m(m-1)}$, die für $n \geq 5$ nach Kendall (18) als Näherungsverteilungen auch für die diskreten Verteilungen der Distanzen in der Gesamtmenge der Rangfolgen angenommen werden können. Erwartungswerte und Standardabweichungen der empirischen Verteilungen und die Werte für $\sigma = \sqrt{\frac{2m+5}{18m(m-1)}}$ (bezeichnet als Ges.S.) finden sich in Tabellen 5.2 und 5.3.

Tabelle 5.2 zeigt, daß die mittleren Distanzen in der Menge der darstellbaren Rangfolgen bei 0,5 liegen. Dies entspricht genau dem Erwartungswert in der Gesamtmenge der Permutationen.

Tabelle 5.3 zeigt, daß auch die Standardabweichungen der Distanzen in der Menge der darstellbaren Rangfolgen in etwa in dem Bereich liegen wie die Standardabweichungen in der Gesamtmenge der Rangfolgen. Sie sind aber generell etwas größer. Dies liegt vor allem an der oben erörterten Tatsache, daß das Auswahlverfahren der Monte-Carlo-Studie nicht jeder der darstellbaren Rangfolgen die gleiche Wahrscheinlichkeit, gezogen zu werden, zuweist. Das wird klar bei der Betrachtung der Standardabweichungen für z.B. 6 Punkte und 5 Dimensionen, 5 Punkte und 4 oder 5 Dimensionen: Da hier alle Rangfolgen zugleich darstellbare Rangfolgen sind, zeigen die Unterschiede der empirisch ermittelten Standardabweichungen von den theoretischen, daß das Auswahlverfahren nicht unter dem Postulat der Uniformverteilung der Rangfolgen stattfand.

Insgesamt lassen Tabellen 5.2 und 5.3 die Annahme zu, daß die darstellbaren Rangfolgen gleichmäßig in der Gesamtmenge der Rangfolgen verstreut sind. Insbesondere sind keine starken Klumpenbildungen von darstellbaren Rangfolgen zu erwarten.

Das benutzte FORTRAN-Programm ist im Anhang unter Programm 1 zu finden.

Tabelle 5.2

ERWARTUNGSWERTE DER NORM. PERMUTATIONSABSTÄNDE
ZWISCHEN DARSTELLBAREN PERMUTATIONEN

PKT	DIM				
	1	2	3	4	5
4	.5010	.4928	.4995	.4973	.5031
5	.5020	.5019	.5060	.4932	.4986
6	.5036	.4982	.4983	.5078	.5004
7	.4993	.4982	.4998	.4988	.5001
8	.4987	.4958	.5058	.4992	.5010
9	.4998	.5001	.5023	.5013	.4996
10	.4979	.5002	.4960	.4973	.4965
11	.4980	.5009	.4980	.5029	.4992
12	.5003	.4998	.5004	.5003	.4992
13	.4970	.5023	.4993	.5006	.5022
14	.5008	.4993	.5022	.4994	.4985
15	.4997	.5006	.4979	.5003	.4989
20	.4979	.5031	.4967	.4982	.5039
25	.5078	.4990	.4955	.5017	.4974
30	.4990	.5022	.5091	.5023	.5029
35	.4977	.4993	.5005	.5041	.5024
40	.4937	.4917	.4983	.4990	.4970

Tabelle 5.3

STANDARDABWEICHUNGEN DER NORM. PERMUTATIONSABSTÄNDE
ZWISCHEN DARSTELLBAREN PERMUTATIONEN

PKT	DIM					
	1	2	3	4	5	GES.S
4	.1861	.1807	.1771	.1807	.1745	.177430
5	.1385	.1365	.1293	.1313	.1273	.124226
6	.1113	.1089	.1047	.1085	.1055	.096225
7	.0934	.0928	.0929	.0847	.0828	.078848
8	.0800	.0810	.0798	.0752	.0759	.066953
9	.0716	.0746	.0697	.0675	.0633	.058267
10	.0612	.0655	.0638	.0602	.0586	.051629
11	.0558	.0614	.0582	.0562	.0530	.046380
12	.0558	.0586	.0551	.0539	.0484	.042121
13	.0495	.0553	.0512	.0494	.0465	.038591
14	.0460	.0517	.0489	.0471	.0437	.035616
15	.0444	.0497	.0471	.0449	.0427	.033073
20	.0280	.0382	.0351	.0346	.0289	.024405
25	.0330	.0315	.0315	.0333	.0248	.019357
30	.0258	.0296	.0304	.0267	.0243	.016046
35	.0219	.0228	.0261	.0224	.0232	.013706
40	.0209	.0226	.0296	.0223	.0214	.011962

5.2 Anzahl der Rangfolgen in Klassen mit gleichem Stress

Mit Hilfe der Risikofunktion soll die Güte von NMDS-Verfahren, die auf Kruskal's Stress fußen, berechnet werden. Dazu muß die Verteilung des Stresses bekannt sein. Die Größe der Klassen von Rangfolgen mit gleichem Stress ist jedoch kombinatorisch nicht auszumachen. Deshalb wurde ein Simulationsverfahren entwickelt, das die Stressverteilung schätzt und das es gleichzeitig ermöglicht, den Stress mit dem Permutationsabstand und Guttman's Alienationskoeffizienten zu vergleichen. Die Punktezahl n wurde variiert mit $n = 4, \dots, 15, 20, 25, 30, 35, 40$, NMDS-Darstellungen wurden jeweils für $p = 1, \dots, 5$ Dimensionen berechnet. Für jedes der Paare (n, p) mit $n \leq 15$ wurden 1000 zufallsgenerierte Rangfolgen untersucht, für $n > 15$ wurden aus Kostengründen weniger als 1000 Rangfolgen untersucht, und zwar

Punktezahl	Anzahl der untersuchten Rangfolgen
4 - 15	1000
20, 25, 30	224
35	194
40	134

Das Programm zur Erzeugung der Matritzen mit Rangplätzen ist im Anhang unter "Programm 2" wiedergegeben.

Für die Zufallsgeneration der Rangfolgen wurde das Programm GGPER aus der Programmbibliothek IMSL (17) benutzt.

Anschließend mußten die NMDS-Lösungen für jede der insgesamt 13.000 Rangfolgen in Räumen der Dimension eins bis fünf gesucht werden. Hierzu wurde das Programm MINISSA (30) so umgestaltet, daß nur noch Kruskal's Streß, Guttman's Alienations-Koeffizient, der Raw Streß von Kruskal und Guttman und zusätz-

lich der normierte Permutations-Abstand s ausgegeben wurden. Das geschah durch

- a. Ausschaltung aller hier überflüssigen Ausgaben auf die Datei Output,
- b. Herausnahme aller hier überflüssigen Unterprogramme zur Einsparung von Rechenzeit und
- c. Einfügung eines neuen Unterprogramms zur Berechnung von s . (vgl. Subroutine "Kentau" in Programm 3)

Die Berechnung der insgesamt 65.000 Werte für jeden der fünf Koeffizienten erforderte einen hohen Aufwand an Rechenzeit. Für die Rechenläufe, die mit dem Programm MINISSA arbeiteten, wurden allein ca. 75.000₁₀ Sekunden CPU-Zeit verbraucht.

Das beschriebene Simulationsverfahren entspricht einem Zufallsexperiment, bei dem θ uniformverteilt ist. Denn jede der möglichen Rangfolgen x hatte die gleiche Wahrscheinlichkeit, generiert zu werden. Da $P_{\theta}(x)$ für alle θ gleich ist, ist θ uniformverteilt.

In Tabelle 5.4 sind die arithmetischen Mittel und Varianzen der empirischen Verteilungen von Kruskal's Streß, dem Permutationsabstand und Guttman's Streß wiedergegeben. Eine Betrachtung der zentralen Momente 3.Ordnung zeigte, daß keines der drei Abstandsmaße normalverteilt ist. Vielmehr ist Kruskal's Streß durchgängig rechts-schief (negative Momente 3.Ordnung), die Permutationsabstände für alle Punkte und Dimensionen links-schief, Guttman's Streß überwiegend rechts-schief verteilt. Die Rechts-Schiefen von Kruskal's Streß-Verteilung erklärt sich aus der Verteilung der darstellbaren Rangfolgen und dem NMDS-Verfahren, das auf Kruskal's Streß gestützt ist. Wegen der recht gleichmäßigen Verteiltheit der darstellbaren Rangfolgen in der Gesamtmenge der Rangfolgen besitzt jede der darstellbaren Rangfolgen einen "Einzugsbereich",

Tabelle 5.4

MITTLERE STRESSWERTE UND VARIANZEN (MAL 1000)

PKT	DIM	KSTR.	PERM.	GSTR.	V.KSTR	V.PERM	V.GSTR
4	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	3	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
4	1	6.4670	173.7746	11.8404	32.4748	100.9563	60.642
5	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
5	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
5	3	0.0000	.0225	0.0000	0.0000	.7031	0.000
5	2	.1348	35.2012	.2462	.5650	59.1469	1.015
5	1	143.1410	178.5373	250.2901	80.4689	70.7987	139.060
6	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.000
6	4	0.0000	.0505	0.0000	0.0000	.7092	0.000
6	3	.0576	14.4586	.0980	.4494	35.0335	.768
6	2	8.0455	107.9838	13.8758	23.5444	70.7303	40.141
6	1	217.0141	217.8263	337.6323	58.4614	56.7234	83.640
7	5	0.0000	.2080	0.0000	0.0000	1.3155	0.000
7	4	.0080	17.6090	.0150	.1093	34.9727	.191
7	3	1.3750	79.9990	2.4470	3.8615	74.0732	6.668
7	2	73.5370	149.0070	120.9030	32.4544	89.4079	51.893
7	1	274.9690	251.4210	401.3670	48.6986	49.6647	67.300
8	5	.0050	14.7110	.0070	.0706	30.2303	.104
8	4	.2720	56.4660	.4820	1.1582	58.1024	2.004
8	3	10.4500	123.4490	17.5460	18.4103	80.7914	30.414
8	2	115.7380	179.6200	177.8100	24.6281	90.1227	36.381
8	1	312.3780	272.7270	437.2680	41.6357	45.3595	58.012
9	5	.2600	52.8060	.4500	1.0146	52.8761	1.722
9	4	4.2920	104.2110	7.2680	6.4574	75.7457	10.773
9	3	56.2130	126.1520	88.3950	18.9934	98.0437	28.769
9	2	146.8100	202.2820	215.7070	21.8288	83.4833	32.590
9	1	340.5430	287.8380	460.6910	36.0839	41.1237	51.401
10	5	1.8860	95.7130	3.1570	3.4965	64.8268	5.798
10	4	14.4170	116.5910	23.4480	16.9362	88.3857	27.262
10	3	82.4970	146.7080	122.9140	15.2424	100.5063	21.880
10	2	171.3590	226.1880	243.1800	19.3295	72.6600	30.074
10	1	365.4580	303.3390	483.2860	30.8094	35.8334	44.632
11	5	10.2370	113.9900	16.6370	9.4892	78.0530	15.333
11	4	48.1740	109.5170	73.2040	13.2461	97.3182	19.244
11	3	100.7530	162.9280	144.5800	12.8061	99.3237	19.046
11	2	192.6270	240.2370	263.4110	17.1767	67.1228	26.881
11	1	383.1080	311.0520	497.9110	27.3523	31.7660	40.491

Tabelle 5.4

FORTSETZUNG: MITTLERE STRESSWERTE UND VARIANZEN (MAL 1000)

PKT	DIM	KSTR.	PERM.	GSTR.	V.KSTR	V.PERM	V.GSTR
12	5	23.5940	120.1600	36.5650	15.3542	87.3410	23.5623
12	4	65.1220	138.0510	95.2280	10.1380	96.1355	14.4174
12	3	116.0760	189.8350	161.8300	11.5929	92.6776	17.5977
12	2	209.6550	256.8200	281.9760	15.1324	56.8062	25.1523
12	1	399.2160	319.2260	509.8040	23.3563	27.8897	35.1481
13	5	43.4850	127.7460	64.4780	9.4785	90.7121	13.5805
13	4	78.5790	160.2800	111.0540	9.2163	95.6994	13.6104
13	3	130.5160	200.0020	177.2440	10.2934	93.2350	16.7725
13	2	223.9240	263.1450	295.1260	13.6628	54.8977	22.8510
13	1	412.0840	325.7350	520.6100	20.3482	25.5465	31.6431
14	5	56.5890	151.8580	80.8070	8.1096	84.3476	11.3621
14	4	90.0010	183.7200	123.2520	8.5911	85.8479	12.7422
14	3	142.3080	217.9990	189.5120	9.4683	86.5726	16.3675
14	2	236.4560	273.8870	307.6660	12.3336	49.7075	21.9705
14	1	424.6690	332.8260	531.5790	18.3727	24.0147	28.4117
15	5	66.9630	165.6140	92.7500	7.2574	82.3887	10.0940
15	4	100.1580	200.7840	134.3900	7.5026	78.8749	11.8729
15	3	152.8060	231.1990	200.4350	8.7673	81.5425	14.9312
15	2	247.6250	284.2040	318.4540	11.1168	39.8091	19.9970
15	1	434.2360	337.2900	539.4050	17.4321	21.6326	27.2740
20	5	109.5179	223.5804	137.9554	12.2449	57.0005	15.3043
20	4	143.2679	250.1339	179.4643	15.4092	58.6683	19.7625
20	3	197.5000	272.4375	245.9330	20.0593	73.3312	25.4342
20	2	292.5848	316.4866	359.9955	22.3850	21.0242	26.8347
20	1	473.0045	359.3527	571.0268	21.2154	19.8170	25.9311
25	5	131.5223	257.8571	159.5893	5.7308	29.9546	7.8433
25	4	165.7098	284.1205	202.1563	6.5070	25.3996	9.2590
25	3	221.0982	301.7188	269.6116	7.5459	52.4118	10.8956
25	2	315.7813	331.5759	383.0536	8.2875	14.5286	12.8089
25	1	493.7768	372.6161	591.8884	9.7048	14.3209	15.7779
30	5	147.2321	281.7232	176.0045	5.1393	12.7080	6.1757
30	4	182.1518	305.2902	219.5268	5.6785	13.7009	7.6492
30	3	237.9598	318.9911	288.6786	6.4268	54.5340	8.5723
30	2	332.6563	347.1205	402.7991	6.5463	11.5280	10.2546
30	1	509.0134	383.0357	608.0625	7.1226	10.0634	11.9515
35	5	158.3299	296.9948	187.8041	4.3398	10.8997	5.2986
35	4	193.0515	318.5412	231.6237	5.0168	10.7352	6.2369
35	3	248.9845	333.6495	300.8763	5.7038	43.2286	7.2922
35	2	343.5825	356.6598	415.3608	5.4957	8.9279	8.0801
35	1	518.3660	390.4381	618.8969	6.2814	9.5786	10.8852
40	5	167.3435	308.7939	197.9008	4.3088	11.8644	5.7644
40	4	202.1985	328.8931	241.9160	4.6664	10.9357	6.3755
40	3	258.4427	345.6794	311.7099	5.2377	32.1294	7.3532
40	2	352.8168	365.3893	426.7786	5.1243	9.2600	8.3788
40	1	526.3511	399.2748	630.4962	5.6527	10.4108	10.7656

der aus all den Rangfolgen besteht, die mit großer Wahrscheinlichkeit durch das NMDS-Verfahren in die im Zentrum des "Einzugsbereichs" liegende darstellbare Rangfolge abgebildet werden. Die Anzahl der Rangfolgen im "Einzugsbereich" nimmt mit wachsendem Abstand (Streß) von der innenliegenden darstellbaren Rangfolge zu, ähnlich, wie die Dichte der Normalverteilung der Abstände innerhalb der Gesamtmenge der Rangfolgen Kendall's tau-Verteilung von links nach rechts zunächst zunehmen. Am Rande des "Einzugsbereichs" beginnen nun die "Einzugsbereiche" anderer darstellbarer Rangfolgen, und die Anzahl der Rangfolgen, die in die erste darstellbare Rangfolge abgebildet werden, nimmt recht rapide ab - ein Hinweis auch auf die Güte des Verfahrens, das die streßminimierende darstellbare Rangfolge bestimmen soll.

Bemerkenswert ist, daß die Permutationsabstände nicht dieser Verteilung folgen, sondern linksschief verteilt sind. Dies ist ein erster Hinweis darauf, daß Streß-Optimalität nicht generell bedeutet, daß auch die Optimalität auf dem ordinalen Meßniveau erzielt wird. Näheres zum Zusammenhang zwischen Streß und Permutationsabstand wird in Kapitel 6 erörtert.

Die empirischen Verteilungen von Kruskal's Streß und des Permutationsabstandes wurden durch Betaverteilungen mit Parametern p und q approximiert. Mit Ausnahme der Konfigurationen mit wenigen Punkten sind diese Approximationen sehr exakt.

Tabelle 5.5 enthält die Parameter p und q der Beta-Verteilungen sowie die R^2 -Werte der Regressionen zwischen den empirischen Verteilungen und den approximierenden Beta-Verteilungen für Kruskal' Streß und den Permutationsabstand. Das Programm zur Erzeugung der Permutationenmatritzen ist im Anhang unter "Programm 2" wiedergegeben. Das modifizierte MINISSA-Programm konnte aus Platzgründen nicht abgedruckt werden.

TABELLE 5.5

KOEFFIZIENTEN DER BETA(P,Q)-VERTEILUNGEN FUER KRUSKALS STRESS
UND DEN PERMUTATIONSABSTAND, APPROXIMATIONSGENAUIGKEITEN

DIM	PKT	KRUSKALS STRESS			PERMUTATIONSABSTAND		
		P	Q	R**2	P	Q	R**2
4	1	.0317	4.8773	.6616	2.2742	10.8128	.3774
5	2	.0568	421.1590	.4849	.3065	8.4015	.8895
5	1	2.5682	15.3734	.9571	5.0454	23.2141	.9954
6	3	.0164	284.1717	.4849	.1534	10.4566	.9825
6	2	.1078	13.2891	.9397	1.9711	16.2828	.9928
6	1	10.5722	38.1446	.9973	11.3167	40.6361	.9993
7	4	.0053	668.6421	.4845	.2314	12.9122	.9815
7	3	.1252	90.9606	.7390	.9931	11.4207	.9907
7	2	4.6830	58.9995	.9911	2.2147	12.6482	.9996
7	1	22.8399	60.2236	.9993	18.9328	56.3705	.9998
8	5	.0050	1002.1278	.7522	.2186	14.6420	.9649
8	4	.0549	201.6591	.6740	.8347	13.9471	.9951
8	3	.3084	29.2009	.9737	1.9231	13.6550	.9994
8	2	19.4129	148.3182	.9989	3.0792	14.0635	.9998
8	1	39.3937	84.5141	.9994	26.0189	69.3838	.9998
9	5	.0654	251.4400	.5931	.8919	15.9978	.9911
9	4	.4356	101.0532	.8370	1.5914	13.6792	.9994
9	3	8.2107	137.8530	.9970	1.3206	9.1475	.9992
9	2	38.4452	223.4252	.9994	4.4812	17.6719	.9995
9	1	58.3953	113.0818	.9996	34.6013	85.6098	.9998
10	5	.2885	152.6881	.7539	1.8755	17.7198	.9987
10	4	.6998	47.8380	.9665	1.4206	10.7639	.9986
10	3	26.7943	297.9966	.9992	1.6714	9.7213	.9994
10	2	64.9523	314.0901	.9996	7.2725	24.8799	.9994
10	1	88.9176	154.3869	.9998	49.6198	113.9589	.9997
11	5	1.1417	110.3822	.9696	1.7757	13.8021	.9987
11	4	12.5413	247.7917	.9981	1.0182	8.2790	.9986
11	3	55.5616	495.9015	.9981	2.0895	10.7351	.9987
11	2	101.3455	424.7775	.9996	9.4921	30.0193	.9995
11	1	120.6388	194.2562	.9998	65.7472	145.6233	.9997

FORTSETZUNG TABELLE 5.5

DIM	PKT	KRUSKALS STRESS			PERMUTATIONSABSTAND		
		P	Q	R**2	P	Q	R**2
12	5	2.2820	94.4367	.9727	1.5451	11.3137	.9984
12	4	38.5099	552.8397	.9982	1.5677	9.7881	.9988
12	3	88.5006	673.9363	.9990	3.2094	13.6967	.9993
12	2	151.4994	571.1135	.9995	14.9333	43.2136	.9994
12	1	175.1206	263.5407	.9998	88.8699	189.5218	.9997
13	5	20.0887	441.8804	.9973	1.6021	10.9392	.9988
13	4	66.9034	784.5122	.9981	2.1952	11.5007	.9992
13	3	139.6580	930.3869	.9989	3.4813	13.9250	.9992
13	2	208.2377	721.7104	.9995	16.6671	46.6711	.9994
13	1	240.7088	343.4167	.9998	109.2960	226.2405	.9996
14	5	45.8808	764.8913	.9968	2.5973	14.5061	.9991
14	4	99.7807	1008.8813	.9980	3.5547	15.7939	.9988
14	3	193.6096	1166.8873	.9988	4.7406	17.0052	.9994
14	2	280.4071	905.4671	.9994	21.7708	57.7174	.9995
14	1	306.9534	415.8529	.9997	127.8174	256.2192	.9996
15	5	79.3671	1105.8704	.9970	3.2059	16.1518	.9986
15	4	160.2665	1439.8699	.9979	4.9782	19.8156	.9990
15	3	257.2018	1425.9903	.9990	5.9492	19.7828	.9994
15	2	373.0573	1133.4841	.9994	36.1983	91.1691	.9995
15	1	350.6308	456.8351	.9988	160.7692	315.8806	.9996
20	5	71.1241	578.3048	.9977	11.7220	40.7065	.9982
20	4	73.9166	442.0158	.9985	13.3807	40.1134	.9986
20	3	77.5968	315.2983	.9990	9.7697	26.0907	.9994
20	2	120.5622	291.4967	.9993	154.5719	333.8275	.9994
20	1	261.4870	291.3344	.9995	210.3021	374.9226	.9996
25	5	457.3009	3019.6829	.9974	54.7366	157.5382	.9988
25	4	540.9024	2723.2520	.9982	89.2918	224.9826	.9991
25	3	668.4785	2354.9675	.9988	22.8390	52.8573	.9989
25	2	993.0763	2151.7466	.9991	347.8217	701.1740	.9935
25	1	1309.9857	1343.0059	.9993	424.3604	714.5071	.9995
30	5	699.7398	4052.8906	.9975	352.7247	899.3011	.9992
30	4	841.3532	3777.6142	.9981	344.6247	784.2183	.9991
30	3	1044.4714	3344.8054	.9984	22.9820	49.0640	.9991
30	2	1722.9212	3456.3621	.9989	591.6037	1112.7142	.9991
30	1	2507.0488	2418.2612	.9990	893.4349	1439.0759	.9991
35	5	1120.1262	5954.5085	.9973	521.6509	1234.7802	.9993
35	4	1194.7263	4993.9141	.9979	599.6805	1282.9033	.9989
35	3	1430.8344	4315.8463	.9985	39.3619	78.6118	.9990
35	2	2565.2992	4901.0275	.9987	1026.3595	1851.3393	.9990
35	1	3279.5074	3047.1178	.9990	1012.3966	1580.5793	.9991
40	5	1255.7785	6248.4178	.9885	467.9149	1047.3836	.9878
40	4	1497.7115	5909.4229	.9866	606.6982	1237.9686	.9870
40	3	1805.2188	5179.7678	.9834	75.3956	142.7128	.9854
40	2	23067.6564	5627.1007	.9755	987.7259	1715.4893	.9830
40	1	4106.1997	3695.0563	.9611	883.1921	1328.7984	.9814

5.3 Anzahl der Rangfolgen mit gleichem Abstand zu einer bestimmten Rangfolge

Die Anzahl der Rangfolgen, die den gleichen Abstand zu einer bestimmten (zugrundeliegenden) Rangfolge haben, sei wie vorher mit $h_{\theta}(i)$ bezeichnet. Zur Unterscheidung der $h_{\theta}(i)$ für Rangfolgen mit unterschiedlichen Anzahlen von Folgengliedern (m) wollen wir statt $h_{\theta}(i)$ $h_{\theta}(i, m)$ schreiben. Nach Kendall (18) kann man $h_{\theta}(i, m)$ rekursiv berechnen:

$$h_{\theta}(i, m) = \sum_{j=a}^b h_{\theta}(j, m - 1)$$

wobei $a = \max [i - m + 1, 0]$

$$b = \min \left[\binom{m-1}{2}, i \right]$$

und $m > 2$.

Die Werte für $n < 7$ (n = Anzahl der Objekte, $m = n \cdot (n - 1) / 2$) stehen in Tabelle 5.6 .

Für große m ($m \geq 8$) ist $S(x, \theta) = \overset{\text{quasi}}{\text{normalverteilt}}$ (vgl. (18)).

TABELLE 5.6

HAEUFIGKEITEN DER PERMUTATIONSABSTAEENDE

		PUNKTEZAHLEN			
PERM.	ABST.	3	4	5	6
0	1	1	1	1	1
1	2	5		9	14
2	2	14		44	104
3	1	29		155	545
4	0	49		440	2260
5	0	71		1068	7889
6	0	90		2298	24087
7	0	101		4489	66013
8	0	101		8095	165425
9	0	90		13640	384320
10	0	71		21670	836604
11	0	49		32683	1720774
12	0	29		47043	3366951
13	0	14		64889	6301715
14	0	5		86054	11333950
15	0	1		110010	19664205
16	0	0		135853	33018631
17	0	0		162337	53808313
18	0	0		187959	85306779
19	0	0		211089	131846699
20	0	0		230131	199019426
21	0	0		243694	293868698
22	0	0		250749	425060810
23	0	0		250749	603012233
24	0	0		243694	839953393
25	0	0		230131	1149906518
26	0	0		211089	1548556267
27	0	0		187959	2052994543
28	0	0		162337	2681325612
29	0	0		135853	3452124397
30	0	0		110010	4383749406
31	0	0		86054	5493521812
32	0	0		64889	6796793172
33	0	0		47043	8305935430
34	0	0		32683	10029297355
35	0	0		21670	11970180517
36	0	0		13640	14125894411
37	0	0		8095	16486953600
38	0	0		4489	19036479131
39	0	0		2298	21749861592
40	0	0		1068	24594733930
41	0	0		440	27531288780
42	0	0		155	30512958144
43	0	0		44	33487453708
44	0	0		9	36398145065
45	0	0		1	39185731988
46	0	0		0	41790147131
47	0	0		0	44152608576
48	0	0		0	46217728808
49	0	0		0	47935579065
50	0	0		0	49263606316
51	0	0		0	50168304699
52	0	0		0	50626553988
53	0	0		0	50626553988
54	0	0		0	50168304699

$h_{\theta}(i, m)$ ist für alle θ gleich. Die Verteilung der Abstände zwischen allen Rangfolgen mit gleicher Anzahl von Folgengliedern (m) berechnet sich deshalb nach der Formel

$$G(i, m) = \sum_{\theta} h_{\theta}(i, m) = \frac{m!}{2} h_{\theta}(i, m)$$

Sei $d(p, m)$ die Anzahl der im R^p darstellbaren Rangfolgen der Größe m , $d(p, m) = \text{card}\{D\}$.

Sei ferner $V(i, p, m) = d(p, m) \cdot h_{\theta}(i, m)$ die Verteilung aller Permutationsabstände zwischen jeweils einem $\theta \in D$ und allen anderen Rangfolgen für alle $\theta \in D$.

$$\text{Mit } W(i, p, m) = \sum_{\theta \in D} h_{\theta}^D(i, m),$$

$$h_{\theta}^D(i, m) = \text{card}\{x \mid S(x, \theta) = i, \quad x \in D, \theta \in D\}$$

sei die Verteilung der Gesamtheit aller Permutationsabstände innerhalb der Klasse der im R^p darstellbaren Rangfolgen bezeichnet.

Dann ist $B(i, p, m) = V(i, p, m) - W(i, p, m)$ die Verteilung aller Abstände zwischen je einem Element θ aus D und einem Element x , das nicht in D liegt, für alle $\theta \in D$ und alle $x \notin D$.

Ferner ist $U(i, p, m) = G(i, m) - V(i, p, m)$ die Verteilung der Abstände zwischen allen θ und x mit $\theta \notin D$ und $x \notin D$.

Schätzungen $\hat{d}(p, m)$ sind nach Kapitel 5.2 möglich, Schätzungen $\hat{W}(i, p, m)$ nach Kapitel 5.1. Daraus ergeben sich $\hat{V}$, $\hat{B}$ und $\hat{U}$.

5.4 Verteilungsannahmen für Rangfolgen

Herkömmliche Testverfahren, die Kendall's tau oder Spearman's rho als Prüfgrößen benutzen, setzen eine Uniformverteilung der gemessenen Rangfolgen voraus. Diese Verteilungsannahme ist für die Zwecke der NMDS nicht geeignet, da zumindest unterstellt werden muß, daß die Wahrscheinlichkeit, eine Rangfolge x zu messen, mit der Distanz von x zur "wahren" zugrundeliegenden Rangfolge θ abnimmt. In Kapitel 4, Beispiel 3, wurde eine Verteilung für X definiert, die für wachsende Permutationsabstände zwischen x und θ linear abnimmt und außerdem die Größen der Mengen $\{x\}_i = \{x \mid S(x, \theta) = i\}$ berücksichtigt:

$$(5.1) \quad P_{\theta}(x) = \frac{i_{\max} + 1 - i}{\sum_{i=0}^{i_{\max}} h_{\theta}(i) (i_{\max} + 1 - i)} \quad *)$$

$$h_{\theta}(i) = \text{card}\{x\}_i$$

$$i_{\max} = \max_{x'} S(x', \theta)$$

$$i = S(x, \theta)$$

Zur Vereinfachung setzen wir in (5.1) für

$$(i_{\max} + 1 - i) = i_m \text{ ein, also}$$

$$(5.2) \quad P_{\theta}(x) = \frac{i_m}{\sum_{i=0}^{i_{\max}} h_{\theta}(i) \cdot i_m}$$

*) Hier sind im Unterschied zu Kap. 4 Wahrscheinlichkeiten einzelner Rangfolgen beschrieben, daher fehlt $h_{\theta}(i)$ im Zähler

Veränderungen innerhalb des linearen Ansatzes sind nur möglich durch Addition von linearen Termen der Art $k = a i + b$ zu i_m . Durch geeignete Wahl von a und b kann $P_{\theta}(x)$ für wachsendes i gleich bleiben oder linear zunehmen.

Schulman (35) entwickelte Verteilungen, die mit konstantem Faktor multiplikativ bzw. logarithmisch-linear abnehmen. Mit den hier benutzten Bezeichnungen stellen sich seine Verteilungen dar als

$$P_{\theta}(x) = \frac{k^i}{\sum_{i=0}^m h_{\theta}(i) \cdot k^i}$$

$$0 \leq k \leq 1.$$

Durch diese Umschreibung läßt sich im übrigen der Rekurs auf einen "Anker" oder Vergleichswert $P_{\theta}(x)$ bei $i = 0$ (der durch k festgelegt ist) vermeiden (vgl. (35)).

Nicht nur in den Naturwissenschaften, sondern auch für Sozialwissenschaftler ist die Normalverteilung attraktiv. Dies wohl nicht nur, weil ihre Annahme die Voraussetzung für einige häufig angewandte statistische Verfahren ist, sondern auch, weil sie eine plausible Beschreibung für meist unbekanntes Fehlerverhalten liefert. Wir wollen deshalb eine für die (nur) diskreten Probleme der Rangfolgenverteilungen nachempfundene Form der Normalverteilung beschreiben und anwenden.

Dabei soll von folgenden Prinzipien ausgegangen werden:

- a. Keine aller möglichen Rangfolgen erhält die Wahrscheinlichkeit Null.
- b. Rangfolgen mit gleichem Permutationsabstand von θ erhalten die gleiche Wahrscheinlichkeit.
- c. Die "Fehlergewichtung" nimmt für wachsenden Permutationsabstand von x zu θ ($S(x, \theta)$) so ab, wie es der rechts vom Erwartungswert liegende Teil der Normalverteilung für stetige X tut. Diese Wahrscheinlichkeitsverteilung soll mit $N^d(\theta, \sigma)$ bezeichnet werden.

Punkt c. beschreibt das Bildungsgesetz für die diskrete Verteilung. Für $n = 3$ Objekte mit $\sigma = k \cdot S_{\max}$,

($S_{\max} = \max_{x'} S(x', \theta)$) ergeben sich folgende Werte für

Gewichte w_i , die aus den Normalverteilungen mit verschiedenen k berechnet wurden:

$S(x, \theta)$	$k =$				
	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$
0	0,398942	0,797884	1,196826	1,595769	1,994711
1	0,377383	0,638896	0,725912	0,656040	0,497386
2	0,319448	0,328020	0,161973	0,045584	0,007711
3	0,241970	0,107982	0,013295	0,000535	0,000007

Die Normierung zur Gesamtsumme eins wird gesichert durch die Berechnungsformel

$$(5.3) \quad P_{\theta}(x=i) = p_i = \frac{w_i}{\sum_i h_{\theta}(i) w_i}$$

Die Prinzipien a. und b. wurden bereits für die oben

angegebenen Verteilungen eingehalten. Aussage b. ist äquivalent zum zweiten von Schulman (35) angegebenen Bildungsgesetz. (Schulmans erste Annahme legt lediglich fest, daß

$$\sum_x P_{\theta}(x) = 1 \text{ ist.}$$

Prinzip a. kann bei Verallgemeinerungen auch aufgehoben werden.

Anstelle der zur Normalverteilung analogen Verteilungen können auch andere bestimmt werden, z.B. solche, die zu anderen kontinuierlichen Verteilungen analog sind. Die Aufgabe, eine dem Fehlerverhalten angemessene Verteilungskonstruktion zu wählen, fällt dem Sozialwissenschaftler beim Planen des Forschungsdesigns zu. Insbesondere würde sich anbieten, eine empirische Verteilung in einem Pretest zu ermitteln. Wir wollen hier das Verhalten der Risikofunktion bei $X \sim N^d(\theta, \sigma)$ für verschiedene k ($\sigma = k \cdot S_{\max}$) zeigen.

5.5. Schätzung der Risiken bei der Verwendung des Permutationsabstandes

Zunächst ist es für die Bestimmung des Risikos wichtig, die Verteilung der Zufallsvariablen zu kennen.

Nach Tabellen 5.2 und 5.3 kann angenommen werden, daß die Verteilung der Distanzen zwischen darstellbaren Rangfolgen sich nicht wesentlich von der Verteilung der Distanzen in der Gesamtheit der Rangfolgen unterscheidet. Die Verteilung der relativen Häufigkeiten der Distanzen für unterschiedliche Punktezahlen n nähert sich mit wachsendem n immer mehr der Dichte einer Normalverteilung mit $\mu = 0,5$ und

$$\sigma_1 = \frac{(2m+5)}{18m(m-1)} \quad ; \quad m=n(n-1)/2$$

(vgl. Kendall, (18))

Wir wollen diese Dichtefunktion auch den Distanzen zwischen den darstellbaren Rangfolgen zugrundelegen.

Die Dichtefunktion der Zufallsvariablen X (das sind die gemessenen Rangfolgen) soll dann durch ihre Asymptote dargestellt werden:

$$(5.4) \quad f_{\theta}(i) = f_{\theta}(x \in \{x_i\}) = \frac{\exp\left(-\frac{i^2}{2\sigma_1^2} - \frac{(i-0.5)^2}{2\sigma_1^2}\right)}{\int_i \exp\left(-\frac{i^2}{2\sigma_1^2} - \frac{(i-0.5)^2}{2\sigma_1^2}\right) di}$$

mit $i = s(x, \theta) \in [0, 1]$.

(5.4) ergibt sich aus (5.3), wenn man anstelle der diskreten Häufigkeitsverteilungen die entsprechenden Normalverteilungen einsetzt sowie die Größen der Mengen $\{x\}_i$ im Zähler berücksichtigt.

σ_2^2 ist die Varianz der Verteilung der X um θ und wird nach Kapitel 5.4 berechnet als

$$\sigma_2^2 = k^2 \cdot s_{\max}^2.$$

Die Verteilung der X um θ drückt "Fehler" bzw. Zufallseinfluß bei der Erhebung der Rangfolgen aus. Um unterschiedliche Fehlersituationen einzubeziehen, werden Risiken für verschiedene k berechnet. Das jeweils zu wählende k ist abhängig vom Forschungsdesign bzw. in einem Pretest zu ermitteln. Die variierenden k geben hier die Möglichkeit, die Robustheit des Verfahrens zu beurteilen. Tabelle 5.7 enthält Werte der Verteilungsfunktionen (5.4)

Als letztes sind für die Konstruktion von Risikofunktionen die Entscheidungsfunktion $\varphi(x)$ und die Verlustfunktion $L(\varphi(x), \theta)$ wichtig. Wann soll bei vorliegendem x eine Entscheidung zugunsten einer darstellbaren Permutation getroffen werden? Das wichtigste Kriterium dafür ist hier der Permutationsabstand zwischen x und der von MINISSA berechneten darstellbaren Rangfolge mit minimalem Streß ($d(x) \in D$). Der maximale Abstand, bei dem die Entscheidung zugunsten der darstellbaren Rangfolge fällt, heiße α . α ist für verschiedene Dimensionen bzw. Punktezahlen sinnvollerweise unterschiedlich zu wählen, weil die zu erwartenden Abstände zwischen x und der Rangfolge der NMDS-Lösung mit zunehmender Dimension und Punktezahl ebenfalls zunehmen. Wir stellen daher Berechnungen für verschiedene α -Werte an.

(5.5) Die Entscheidungsfunktion $\varphi(x)$ sei definiert durch

$$\varphi(x) = \begin{cases} d(x) \in D & \text{falls } s(d(x), x) \leq \alpha \\ x & \text{sonst} \end{cases}$$

Die Verlustfunktion ist definiert durch

$$L(\varphi(x), \theta) = s(\varphi(x), \theta) \quad (\text{normierter Permutationsabstand})$$

Tabelle 5.7

KOEFFIZIENT ZUM BESTIMMUNG DER FEHLEHVARIANZ K=1.0000

PKT	/	x 0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
4		0.0000	.0111	.0490	.1430	.3124	.5344	.7458	.8922	.9659	.9928	1.0000
5		0.0000	.0007	.0088	.0593	.2267	.5246	.8086	.9539	.9937	.9995	1.0000
6		0.0000	.0000	.0010	.0207	.1596	.5191	.8626	.9837	.9993	1.0000	1.0000
7		0.0000	.0000	.0001	.0061	.1088	.5157	.9052	.9951	.9999	1.0000	1.0000
8		0.0000	.0000	.0000	.0015	.0717	.5133	.9370	.9988	1.0000	1.0000	1.0000
9		0.0000	.0000	.0000	.0003	.0455	.5116	.9596	.9997	1.0000	1.0000	1.0000
10		0.0000	.0000	.0000	.0001	.0278	.5103	.9753	1.0000	1.0000	1.0000	1.0000
11		0.0000	.0000	.0000	.0000	.0164	.5092	.9854	1.0000	1.0000	1.0000	1.0000
12		0.0000	.0000	.0000	.0000	.0093	.5084	.9917	1.0000	1.0000	1.0000	1.0000
13		0.0000	.0000	.0000	.0000	.0050	.5077	.9955	1.0000	1.0000	1.0000	1.0000
14		0.0000	.0000	.0000	.0000	.0026	.5071	.9977	1.0000	1.0000	1.0000	1.0000
15		0.0000	.0000	.0000	.0000	.0013	.5066	.9988	1.0000	1.0000	1.0000	1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5049	1.0000	1.0000	1.0000	1.0000	1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5039	1.0000	1.0000	1.0000	1.0000	1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5032	1.0000	1.0000	1.0000	1.0000	1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5027	1.0000	1.0000	1.0000	1.0000	1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5024	1.0000	1.0000	1.0000	1.0000	1.0000

KOEFFIZIENT ZUM BESTIMMUNG DER FEHLEHVARIANZ K=.9000

PKT	/	x 0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
4		0.0000	.0115	.0504	.1466	.3188	.5423	.7527	.8965	.9677	.9933	1.0000
5		0.0000	.0007	.0091	.0608	.2306	.5303	.8129	.9556	.9940	.9996	1.0000
6		0.0000	.0000	.0010	.0211	.1620	.5235	.8653	.9842	.9993	1.0000	1.0000
7		0.0000	.0000	.0001	.0062	.1104	.5193	.9069	.9953	.9999	1.0000	1.0000
8		0.0000	.0000	.0000	.0016	.0726	.5164	.9381	.9988	1.0000	1.0000	1.0000
9		0.0000	.0000	.0000	.0003	.0461	.5143	.9604	.9997	1.0000	1.0000	1.0000
10		0.0000	.0000	.0000	.0001	.0282	.5127	.9757	1.0000	1.0000	1.0000	1.0000
11		0.0000	.0000	.0000	.0000	.0166	.5114	.9856	1.0000	1.0000	1.0000	1.0000
12		0.0000	.0000	.0000	.0000	.0094	.5104	.9919	1.0000	1.0000	1.0000	1.0000
13		0.0000	.0000	.0000	.0000	.0051	.5095	.9956	1.0000	1.0000	1.0000	1.0000
14		0.0000	.0000	.0000	.0000	.0027	.5088	.9977	1.0000	1.0000	1.0000	1.0000
15		0.0000	.0000	.0000	.0000	.0013	.5081	.9988	1.0000	1.0000	1.0000	1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5066	1.0000	1.0000	1.0000	1.0000	1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5048	1.0000	1.0000	1.0000	1.0000	1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5040	1.0000	1.0000	1.0000	1.0000	1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5034	1.0000	1.0000	1.0000	1.0000	1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5029	1.0000	1.0000	1.0000	1.0000	1.0000

KOEFFIZIENT ZUM BESTIMMUNG DER FEHLEHVARIANZ K=.8000

PKT	/	x 0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
4		0.0000	.0119	.0524	.1517	.3277	.5532	.7622	.9023	.9702	.9940	1.0000
5		0.0000	.0007	.0099	.0626	.2361	.5322	.8187	.9578	.9945	.9996	1.0000
6		0.0000	.0000	.0011	.0218	.1655	.5297	.8689	.9849	.9993	1.0000	1.0000
7		0.0000	.0000	.0001	.0064	.1125	.5244	.9092	.9955	.9999	1.0000	1.0000
8		0.0000	.0000	.0000	.0016	.0740	.5208	.9395	.9989	1.0000	1.0000	1.0000
9		0.0000	.0000	.0000	.0003	.0469	.5181	.9613	.9998	1.0000	1.0000	1.0000
10		0.0000	.0000	.0000	.0001	.0287	.5161	.9762	1.0000	1.0000	1.0000	1.0000
11		0.0000	.0000	.0000	.0000	.0169	.5144	.9859	1.0000	1.0000	1.0000	1.0000
12		0.0000	.0000	.0000	.0000	.0095	.5131	.9920	1.0000	1.0000	1.0000	1.0000
13		0.0000	.0000	.0000	.0000	.0052	.5120	.9957	1.0000	1.0000	1.0000	1.0000
14		0.0000	.0000	.0000	.0000	.0027	.5111	.9977	1.0000	1.0000	1.0000	1.0000
15		0.0000	.0000	.0000	.0000	.0013	.5103	.9989	1.0000	1.0000	1.0000	1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5076	1.0000	1.0000	1.0000	1.0000	1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5060	1.0000	1.0000	1.0000	1.0000	1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5050	1.0000	1.0000	1.0000	1.0000	1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5043	1.0000	1.0000	1.0000	1.0000	1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5037	1.0000	1.0000	1.0000	1.0000	1.0000

Tabelle 5.7. Fortsetzung

KOEFFIZIENT ZUM BESTIMMUNG DER FEHLEKVARIANZ K= .7000

PKT	/	X									
		0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90 1.00
4		0.0000	.0127	.0554	.1592	.3407	.5089	.7755	.9103	.9735	.9948 1.0000
5		0.0000	.0008	.0099	.0654	.2442	.5497	.8270	.9608	.9950	.9997 1.0000
6		0.0000	.0000	.0011	.0227	.1706	.5387	.8742	.9659	.9994	1.0000 1.0000
7		0.0000	.0000	.0001	.0067	.1158	.5319	.9125	.9958	1.0000	1.0000 1.0000
8		0.0000	.0000	.0000	.0017	.0760	.5271	.9416	.9989	1.0000	1.0000 1.0000
9		0.0000	.0000	.0000	.0004	.0482	.5236	.9626	.9998	1.0000	1.0000 1.0000
10		0.0000	.0000	.0000	.0001	.0294	.5210	.9770	1.0000	1.0000	1.0000 1.0000
11		0.0000	.0000	.0000	.0000	.0173	.5188	.9864	1.0000	1.0000	1.0000 1.0000
12		0.0000	.0000	.0000	.0000	.0098	.5171	.9923	1.0000	1.0000	1.0000 1.0000
13		0.0000	.0000	.0000	.0000	.0053	.5157	.9958	1.0000	1.0000	1.0000 1.0000
14		0.0000	.0000	.0000	.0000	.0028	.5145	.9978	1.0000	1.0000	1.0000 1.0000
15		0.0000	.0000	.0000	.0000	.0014	.5134	.9989	1.0000	1.0000	1.0000 1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5099	1.0000	1.0000	1.0000	1.0000 1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5079	1.0000	1.0000	1.0000	1.0000 1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5065	1.0000	1.0000	1.0000	1.0000 1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5056	1.0000	1.0000	1.0000	1.0000 1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5049	1.0000	1.0000	1.0000	1.0000 1.0000

KOEFFIZIENT ZUM BESTIMMUNG DER FEHLEKVARIANZ K= .6000

PKT	/	X									
		0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90 1.00
4		0.0000	.0138	.0602	.1711	.3507	.5524	.7949	.9214	.9779	.9959 1.0000
5		0.0000	.0009	.0108	.0700	.2588	.5571	.8392	.9651	.9958	.9997 1.0000
6		0.0000	.0000	.0012	.0242	.1786	.5525	.8614	.9874	.9995	1.0000 1.0000
7		0.0000	.0000	.0001	.0071	.1209	.5432	.9174	.9962	1.0000	1.0000 1.0000
8		0.0000	.0000	.0000	.0018	.0792	.5368	.9447	.9990	1.0000	1.0000 1.0000
9		0.0000	.0000	.0000	.0004	.0501	.5321	.9645	.9998	1.0000	1.0000 1.0000
10		0.0000	.0000	.0000	.0001	.0306	.5285	.9761	1.0000	1.0000	1.0000 1.0000
11		0.0000	.0000	.0000	.0000	.0179	.5256	.9870	1.0000	1.0000	1.0000 1.0000
12		0.0000	.0000	.0000	.0000	.0101	.5233	.9926	1.0000	1.0000	1.0000 1.0000
13		0.0000	.0000	.0000	.0000	.0055	.5213	.9960	1.0000	1.0000	1.0000 1.0000
14		0.0000	.0000	.0000	.0000	.0029	.5197	.9979	1.0000	1.0000	1.0000 1.0000
15		0.0000	.0000	.0000	.0000	.0014	.5183	.9989	1.0000	1.0000	1.0000 1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5135	1.0000	1.0000	1.0000	1.0000 1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5107	1.0000	1.0000	1.0000	1.0000 1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5089	1.0000	1.0000	1.0000	1.0000 1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5076	1.0000	1.0000	1.0000	1.0000 1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5066	1.0000	1.0000	1.0000	1.0000 1.0000

KOEFFIZIENT ZUM BESTIMMUNG DER FEHLEKVARIANZ K= .5000

PKT	/	X									
		0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90 1.00
4		0.0000	.0159	.0685	.1913	.3938	.6298	.8241	.9372	.9837	.9972 1.0000
5		0.0000	.0010	.0123	.0781	.2781	.5753	.8578	.9713	.9968	.9998 1.0000
6		0.0000	.0000	.0014	.0270	.1923	.5749	.8939	.9894	.9996	1.0000 1.0000
7		0.0000	.0000	.0001	.0079	.1296	.5619	.9250	.9968	1.0000	1.0000 1.0000
8		0.0000	.0000	.0000	.0020	.0847	.5528	.9495	.9992	1.0000	1.0000 1.0000
9		0.0000	.0000	.0000	.0004	.0535	.5461	.9674	.9998	1.0000	1.0000 1.0000
10		0.0000	.0000	.0000	.0001	.0326	.5409	.9798	1.0000	1.0000	1.0000 1.0000
11		0.0000	.0000	.0000	.0000	.0191	.5368	.9880	1.0000	1.0000	1.0000 1.0000
12		0.0000	.0000	.0000	.0000	.0108	.5334	.9932	1.0000	1.0000	1.0000 1.0000
13		0.0000	.0000	.0000	.0000	.0058	.5307	.9963	1.0000	1.0000	1.0000 1.0000
14		0.0000	.0000	.0000	.0000	.0030	.5283	.9980	1.0000	1.0000	1.0000 1.0000
15		0.0000	.0000	.0000	.0000	.0015	.5263	.9990	1.0000	1.0000	1.0000 1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5194	1.0000	1.0000	1.0000	1.0000 1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5154	1.0000	1.0000	1.0000	1.0000 1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5128	1.0000	1.0000	1.0000	1.0000 1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5109	1.0000	1.0000	1.0000	1.0000 1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5095	1.0000	1.0000	1.0000	1.0000 1.0000

Tabelle 5.7, Fortsetzung

KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K = .4000$

PKT	/	X									
		0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90 1.00
4		0.0000	.0201	.0651	.2300	.4537	.7924	.8689	.9590	.9909	.9987 1.0000
5		0.0000	.0013	.0154	.0942	.3184	.6446	.8876	.9801	.9961	.9999 1.0000
6		0.0000	.0000	.0018	.0325	.2187	.6150	.9133	.9925	.9998	1.0000 1.0000
7		0.0000	.0000	.0001	.0095	.1466	.5955	.9375	.9977	1.0000	1.0000 1.0000
8		0.0000	.0000	.0000	.0024	.0954	.5617	.9573	.9994	1.0000	1.0000 1.0000
9		0.0000	.0000	.0000	.0005	.0601	.5715	.9722	.9999	1.0000	1.0000 1.0000
10		0.0000	.0000	.0000	.0000	.0001	.0365	.5636	.9827	1.0000	1.0000 1.0000
11		0.0000	.0000	.0000	.0000	.0014	.5572	.9897	1.0000	1.0000	1.0000 1.0000
12		0.0000	.0000	.0000	.0000	.0120	.5521	.9941	1.0000	1.0000	1.0000 1.0000
13		0.0000	.0000	.0000	.0000	.0065	.5478	.9968	1.0000	1.0000	1.0000 1.0000
14		0.0000	.0000	.0000	.0000	.0034	.5441	.9983	1.0000	1.0000	1.0000 1.0000
15		0.0000	.0000	.0000	.0000	.0017	.5410	.9991	1.0000	1.0000	1.0000 1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5303	1.0000	1.0000	1.0000	1.0000 1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5241	1.0000	1.0000	1.0000	1.0000 1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5200	1.0000	1.0000	1.0000	1.0000 1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5171	1.0000	1.0000	1.0000	1.0000 1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5149	1.0000	1.0000	1.0000	1.0000 1.0000

KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K = .3000$

PKT	/	X									
		0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90 1.00
4		0.0000	.0309	.1255	.3171	.5735	.8004	.9331	.9844	.9975	.9998 1.0000
5		0.0000	.0021	.0240	.1345	.4076	.7381	.9343	.9913	.9994	1.0000 1.0000
6		0.0000	.0001	.0028	.0471	.2602	.6946	.9452	.9964	.9999	1.0000 1.0000
7		0.0000	.0000	.0002	.0139	.1874	.6641	.9586	.9988	1.0000	1.0000 1.0000
8		0.0000	.0000	.0000	.0035	.1215	.6417	.9708	.9997	1.0000	1.0000 1.0000
9		0.0000	.0000	.0000	.0007	.0763	.6247	.9806	.9999	1.0000	1.0000 1.0000
10		0.0000	.0000	.0000	.0001	.0462	.6113	.9877	1.0000	1.0000	1.0000 1.0000
11		0.0000	.0000	.0000	.0000	.0270	.6005	.9926	1.0000	1.0000	1.0000 1.0000
12		0.0000	.0000	.0000	.0000	.0152	.5916	.9957	1.0000	1.0000	1.0000 1.0000
13		0.0000	.0000	.0000	.0000	.0082	.5842	.9976	1.0000	1.0000	1.0000 1.0000
14		0.0000	.0000	.0000	.0000	.0043	.5779	.9988	1.0000	1.0000	1.0000 1.0000
15		0.0000	.0000	.0000	.0000	.0021	.5725	.9994	1.0000	1.0000	1.0000 1.0000
20		0.0000	.0000	.0000	.0000	.0000	.5537	1.0000	1.0000	1.0000	1.0000 1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5427	1.0000	1.0000	1.0000	1.0000 1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5355	1.0000	1.0000	1.0000	1.0000 1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5303	1.0000	1.0000	1.0000	1.0000 1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5265	1.0000	1.0000	1.0000	1.0000 1.0000

KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K = .2000$

PKT	/	X									
		0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90 1.00
4		0.0000	.0715	.2609	.5527	.8142	.9506	.9919	.9992	1.0000	1.0000 1.0000
5		0.0000	.0064	.0635	.2620	.6447	.9064	.9883	.9993	1.0000	1.0000 1.0000
6		0.0000	.0002	.0088	.1107	.4723	.8606	.9874	.9997	1.0000	1.0000 1.0000
7		0.0000	.0000	.0008	.0352	.3277	.8204	.9887	.9999	1.0000	1.0000 1.0000
8		0.0000	.0000	.0000	.0092	.2173	.7863	.9911	1.0000	1.0000	1.0000 1.0000
9		0.0000	.0000	.0000	.0020	.1382	.7578	.9936	1.0000	1.0000	1.0000 1.0000
10		0.0000	.0000	.0000	.0004	.0845	.7340	.9957	1.0000	1.0000	1.0000 1.0000
11		0.0000	.0000	.0000	.0001	.0496	.7139	.9973	1.0000	1.0000	1.0000 1.0000
12		0.0000	.0000	.0000	.0000	.0280	.6968	.9984	1.0000	1.0000	1.0000 1.0000
13		0.0000	.0000	.0000	.0000	.0152	.6821	.9991	1.0000	1.0000	1.0000 1.0000
14		0.0000	.0000	.0000	.0000	.0079	.6694	.9995	1.0000	1.0000	1.0000 1.0000
15		0.0000	.0000	.0000	.0000	.0039	.6583	.9997	1.0000	1.0000	1.0000 1.0000
20		0.0000	.0000	.0000	.0000	.0001	.6190	1.0000	1.0000	1.0000	1.0000 1.0000
25		0.0000	.0000	.0000	.0000	.0000	.5952	1.0000	1.0000	1.0000	1.0000 1.0000
30		0.0000	.0000	.0000	.0000	.0000	.5792	1.0000	1.0000	1.0000	1.0000 1.0000
35		0.0000	.0000	.0000	.0000	.0000	.5679	1.0000	1.0000	1.0000	1.0000 1.0000
40		0.0000	.0000	.0000	.0000	.0000	.5593	1.0000	1.0000	1.0000	1.0000 1.0000

Die Risikofunktion $R(\varphi, \theta)$ wird nun folgendermaßen berechnet:

Fall 1: $\theta \in D$, d.h. θ ist eine darstellbare Rangfolge. Dann tritt genau dann ein Verlust ein, wenn $s(x, \theta) > \alpha$ ist, vorausgesetzt, im Gebiet mit Radius α um θ ist θ die einzige darstellbare Rangfolge. Diese Voraussetzung trifft aber wegen der Verteilung der Abstände zwischen darstellbaren Rangfolgen bei entsprechend kleinem α zu.

(5.6) sei $\theta \in D$, dann wird $\hat{R}_1(\varphi, \theta)$ geschätzt nach

$$\hat{R}_1(\varphi, \theta) = \int_{\alpha}^1 f_{\theta}(i) \cdot s(\varphi(x), \theta) \, di = \int_{\alpha}^1 f_{\theta}(i) \cdot i \, di =$$

$$\frac{\int_{\alpha}^1 \exp\left(-\frac{i^2 \sigma_1^2 + (i-0,5)^2 \sigma_2^2}{2\sigma_1^2 - \sigma_2^2}\right) \cdot i \, di}{\int_0^1 \exp\left(-\frac{i^2 \sigma_1^2 + (i-0,5)^2 \sigma_2^2}{2\sigma_1^2 \cdot \sigma_2^2}\right) \, di}$$

Fall 2: Sei $\theta \notin D$. Dann entsteht ein Verlust, falls $s(d(x), x) \leq \alpha$ ist bzw. x in einem der Annahmebereiche A_j der Hypothesen " θ ist gleich $d_j \in D$ " fällt. Man gehe davon aus, daß für alle j alle $x_j \in A_j$ im Durchschnitt über alle θ normal verteilt um θ sind. Für $p < n-3$ wird dann $\hat{R}_2(\varphi, \theta)$ folgendermaßen berechnet:
Sei $\theta \notin D$

$$\hat{R}_2(\varphi, \theta) = \int_0^1 f_{\theta}(i) \cdot i \cdot \hat{P}(x \in \bigcup_j A_j) \cdot di =$$

$$\frac{\hat{P}(x \in \bigcup_j A_j) \cdot \int_0^1 \exp\left(-\frac{i^2 \sigma_1^2 + (i-0,5)^2 \sigma_2^2}{2\sigma_1^2 \cdot \sigma_2^2}\right) \cdot i \cdot di}{\int_0^1 \exp\left(-\frac{i^2 \sigma_1^2 + (i-0,5)^2 \sigma_2^2}{2\sigma_1^2 \cdot \sigma_2^2}\right) \, di}$$

$\hat{P}(x \in \bigcup_j A_j)$ ist die Schätzung für die Wahrscheinlichkeit, daß X bei einem Zufallsexperiment in einen der

Annahmebereiche A_j fällt. Diese Konstante wird geschätzt aus der Approximation der empirisch ermittelten Verteilung der $s(d(x), x)$, vgl. Tabelle 5.5.

Für die Schätzung des Bayes-Risikos $\hat{R}(\psi, H)$ muß nun nur noch die Wahrscheinlichkeit dafür, daß θ aus D ist, also $H\{\theta \mid \theta \in D\} = H(D)$ bestimmt werden. $\hat{H}(D)$ kann ebenfalls aus der Verteilung von $s(d(x), x)$, vgl. Tabelle 5.5, bestimmt werden. Dann ist

$$(5.7) \quad \hat{R}(\psi, H) = \hat{H}(D) \hat{R}_1(\psi, \theta) + (1 - \hat{H}(D)) \hat{R}_2(\psi, \theta)$$

Tabelle 5.8 zeigt Werte von $\hat{R}(\psi, H)$ für verschiedene Punktezahlen n , Dimensionen p , Entscheidungskriterien α und Fehlerkoeffizienten k . Die $\hat{R}(\psi, H)$ können als zu erwartende normierte Permutationsabstände "Verluste" zwischen der zugrundeliegenden "wahren" Rangfolge θ und der Rangfolge $\psi(x)$, auf die die Entscheidung fällt, interpretiert werden. Das zugehörige Programm findet sich im Anhang unter "Programm 3".

$\hat{R}(\psi, H)$ ist recht robust gegenüber Veränderungen des Koeffizienten k . Wenn wir von einem starken Fehler mit $k = 1,0$ ausgehen, so können folgende Maximen aufgestellt werden für die Anzahl der Punkte, die notwendig sind, um eine Skalierung in Räumen der Dimension p mit geringem Risiko (unter 0.01) und $\alpha \leq 0,15$ zu erlauben:

Tabelle 5.9:

Dimensionen	notwendige Mindestzahl von Punkten
1	7
2	9
3	12
4	14
5	über 15

Dabei ist zu beachten, daß $\hat{R}(\psi, H) = 0,01$ bei $n=10$ Punkten einen zu erwartenden Verlust von $m(m-1) \cdot 0,01 = 19,8$ Permutationsschritten bedeutet.

TABELLE 5.8

BAYES-RISIKEN, AUF GRUNDLAGE DER PERMUTATIONSABSTÄNDE GESCHÄTZT

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K=1.0000$)

PKT	DIN	ALPHA							
		0.00	.01	.05	.10	.15	.20	.25	.30
4	1	.0346	.0667	.1931	.2878	.3510	.3945	.4244	.4443
5	1	.0000	.0000	.0057	.0665	.1896	.3170	.4061	.4582
	2	.1780	.2916	.4036	.4443	.4654	.4766	.4819	.4823
6	1	.0000	.0000	.0000	.0022	.0422	.1775	.3457	.4493
	2	.0002	.0003	.0427	.1790	.3161	.4089	.4590	.4819
	3	.2894	.4097	.4573	.4740	.4822	.4869	.4894	.4888
7	1	.0000	.0000	.0000	.0000	.0044	.0623	.2318	.4034
	2	.0000	.0000	.0001	.0109	.0852	.2326	.3735	.4548
	3	.0033	.0066	.0940	.2266	.3320	.4033	.4472	.4722
	4	.2580	.3821	.4523	.4757	.4859	.4911	.4937	.4944
8	1	.0000	.0000	.0000	.0000	.0004	.0179	.1388	.3470
	2	.0000	.0000	.0000	.0010	.0288	.1528	.3315	.4479
	3	.0000	.0000	.0008	.0339	.1557	.3133	.4244	.4758
	4	.0217	.0424	.1889	.3140	.3921	.4391	.4665	.4818
	5	.2681	.3918	.4606	.4818	.4903	.4942	.4961	.4968
9	1	.0000	.0000	.0000	.0000	.0000	.0039	.0738	.2888
	2	.0000	.0000	.0000	.0001	.0094	.0993	.2952	.4425
	3	.0000	.0000	.0000	.0069	.0633	.1940	.3394	.4367
	4	.0000	.0000	.0100	.0889	.2224	.3459	.4277	.4704
	5	.0306	.0594	.2353	.3635	.4317	.4667	.4840	.4922
10	1	.0000	.0000	.0000	.0000	.0000	.0003	.0238	.2077
	2	.0000	.0000	.0000	.0000	.0012	.0451	.2420	.4361
	3	.0000	.0000	.0000	.0011	.0286	.1456	.3170	.4385
	4	.0000	.0000	.0002	.0161	.1031	.2535	.3865	.4604
	5	.0000	.0000	.0195	.1425	.3026	.4144	.4693	.4902
11	1	.0000	.0000	.0000	.0000	.0000	.0000	.0079	.1562
	2	.0000	.0000	.0000	.0000	.0001	.0159	.1840	.4255
	3	.0000	.0000	.0000	.0001	.0121	.1082	.2995	.4415
	4	.0000	.0000	.0000	.0053	.0540	.1775	.3246	.4287
	5	.0000	.0000	.0008	.0358	.1654	.3275	.4354	.4815

FORTSETZUNG TABELLE 5.8

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K=1.0000$)[illegible]

FORTSETZUNG TABELLE 5.8

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K = .5000$)

PKT	DIM	ALPHA							
		.0.00	.01	.05	.10	.15	.20	.25	.30
4	1	.0345	.0666	.1928	.2875	.3508	.3945	.4249	.4456
5	1	.0000	.0000	.0057	.0665	.1896	.3170	.4081	.4582
	2	.1780	.2916	.4006	.4443	.4655	.4771	.4837	.4875
6	1	.0000	.0000	.0000	.0022	.0422	.1775	.3457	.4493
	2	.0002	.0003	.0427	.1790	.3161	.4089	.4590	.4819
	3	.2894	.4097	.4573	.4740	.4822	.4870	.4901	.4920
7	1	.0000	.0000	.0000	.0000	.0044	.0623	.2318	.4034
	2	.0000	.0000	.0001	.0109	.0852	.2326	.3735	.4548
	3	.0033	.0066	.0940	.2266	.3320	.4033	.4472	.4722
	4	.2580	.3821	.4523	.4757	.4859	.4911	.4938	.4953
8	1	.0000	.0000	.0000	.0000	.0004	.0179	.1388	.3470
	2	.0000	.0000	.0000	.0010	.0238	.1528	.3315	.4479
	3	.0000	.0000	.0008	.0339	.1557	.3133	.4244	.4758
	4	.0217	.0424	.1889	.3140	.3921	.4391	.4665	.4818
	5	.2681	.3916	.4606	.4818	.4903	.4942	.4961	.4970
9	1	.0000	.0000	.0000	.0000	.0000	.0039	.0736	.2888
	2	.0000	.0000	.0000	.0001	.0094	.0993	.2952	.4425
	3	.0000	.0000	.0000	.0069	.0633	.1940	.3394	.4367
	4	.0000	.0000	.0100	.0889	.2224	.3459	.4277	.4704
	5	.0306	.0594	.2353	.3635	.4317	.4667	.4840	.4922
10	1	.0000	.0000	.0000	.0000	.0000	.0003	.0238	.2077
	2	.0000	.0000	.0000	.0000	.0012	.0451	.2420	.4361
	3	.0000	.0000	.0000	.0011	.0286	.1456	.3179	.4385
	4	.0000	.0000	.0002	.0161	.1031	.2535	.3865	.4604
	5	.0000	.0000	.0195	.1425	.3026	.4144	.4693	.4902
11	1	.0000	.0000	.0000	.0000	.0000	.0000	.0079	.1562
	2	.0000	.0000	.0000	.0000	.0001	.0159	.1840	.4255
	3	.0000	.0000	.0000	.0001	.0121	.1082	.2995	.4415
	4	.0000	.0000	.0000	.0053	.0540	.1775	.3246	.4267
	5	.0000	.0000	.0008	.0358	.1654	.3275	.4354	.4815

[illegible]

FORTSETZUNG TABELLE 5.8

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K = .1000$)

PKT	DIM	ALPHA							
		.0.00	.0.01	.0.05	.10	.15	.20	.25	.30
4	1	.0345	.0566	.1928	.2875	.3508	.3945	.4250	.4459
5	1	.0000	.0000	.0057	.0665	.1894	.3167	.4077	.4578
	2	.1778	.2915	.4004	.4441	.4652	.4768	.4835	.4873
6	1	.0000	.0000	.0000	.0022	.0420	.1768	.3442	.4474
	2	.0002	.0003	.0425	.1782	.3148	.4071	.4570	.4798
	3	.2681	.4085	.4556	.4731	.4817	.4861	.4891	.4911
7	1	.0000	.0000	.0000	.0000	.0044	.0618	.2301	.4006
	2	.0000	.0000	.0001	.0108	.0846	.2309	.3709	.4516
	3	.0033	.0066	.0933	.2210	.3296	.4005	.4440	.4688
	4	.2562	.3796	.4491	.4736	.4853	.4889	.4918	.4935
8	1	.0000	.0000	.0000	.0000	.0003	.0178	.1376	.3440
	2	.0000	.0000	.0000	.0009	.0285	.1514	.3286	.4440
	3	.0000	.0000	.0008	.0336	.1543	.3105	.4207	.4716
	4	.0215	.0420	.1872	.3114	.3890	.4353	.4625	.4777
	5	.2658	.3884	.4563	.4780	.4901	.4910	.4935	.4945
9	1	.0000	.0000	.0000	.0000	.0000	.0039	.0731	.2859
	2	.0000	.0000	.0000	.0001	.0093	.0983	.2923	.4381
	3	.0000	.0000	.0000	.0069	.0626	.1921	.3360	.4323
	4	.0000	.0000	.0099	.0880	.2202	.3425	.4234	.4658
	5	.0303	.0588	.2329	.3600	.4280	.4622	.4795	.4875
10	1	.0000	.0000	.0000	.0000	.0000	.0003	.0236	.2055
	2	.0000	.0000	.0000	.0000	.0012	.0447	.2394	.4314
	3	.0000	.0000	.0000	.0011	.0283	.1441	.3145	.4339
	4	.0000	.0000	.0002	.0159	.1020	.2508	.3824	.4555
	5	.0000	.0000	.0193	.1409	.2994	.4100	.4644	.4850
11	1	.0000	.0000	.0000	.0000	.0000	.0000	.0078	.1545
	2	.0000	.0000	.0000	.0000	.0001	.0157	.1820	.4207
	3	.0000	.0000	.0000	.0001	.0120	.1070	.2962	.4366
	4	.0000	.0000	.0000	.0052	.0534	.1755	.3210	.4240
	5	.0000	.0000	.0008	.0354	.1635	.3239	.4306	.4762

Fehlermöglichkeiten:

- a. Die Schätzungen von $\hat{P}(x \in \bigcup_j A_j)$ und $\hat{H}(D)$ sind nach dem Satz von Moivre-Laplace asymptotisch normalverteilt mit $N(p, \frac{p(1-p)}{n})$, wobei p der wahre, zu schätzende Wert ist. Tabellen mit Konfidenzintervallen für Schätzungen aus Stichproben unterschiedlicher Größe sind z.B. bei Hald (49) angegeben. Bei einer Stichprobe von 1000 sind die Schätzungen sehr präzise. Das größte 95%-Konfidenzintervall hat der Schätzwert $p=0.5$ mit $[0.516, 0.486]$.
- b. Die Fehler, die durch die Verwendung der kontinuierlichen Dichtefunktionen entstehen, sind schon für kleine n ($n > 5$) zu vernachlässigen, da die Anzahl der Rangfolgen mit wachsendem n schnell sehr groß sind.
- c. Die Verwendung der Routine MINISSA bei der Monte-Carlo-Studie bewirkte bei kleinen Punktezahlen^{und} im Verhältnis großen Anzahlen von Dimensionen eine Überschätzung des Anteils der darstellbaren Rangfolgen, da eine Rangfolge häufig mit beliebig kleinem Streß dargestellt werden kann, ohne daß sie zur Menge der darstellbaren Rangfolgen gehört. Es wird daher das Risiko nur für die Fälle $p \leq n-3$ berechnet.
- d. Ferner ist festzustellen, daß bei der Berechnung von $\hat{R}_2(\varphi, \theta)$ die Multiplikation mit der Konstanten $\hat{P}(x \in \bigcup_j A_j)$ sowie die Tatsache, daß $\hat{R}_2(\varphi, \theta)$ für alle $\theta \in D$ identisch ist, nur dann gültig ist, wenn der Anteil der darstellbaren Rangfolgen an der Gesamtmenge der Rangfolgen klein sowie die Gesamtmenge von Rangfolgen groß ist. Dadurch ist das $\hat{R}_2(\varphi, H)$ für kleine n , z.B. $n < 6$ und großem Anteil von darstellbaren Rangfolgen (z.B. $p > n-3$) mit Fehler behaftet. Da NMDS-Darstellungen bei so kleinen Punktezahlen bzw. in Räumen, deren Dimension groß ist im Verhältnis zur Anzahl der Punkte, ohnehin wegen eines hohen Risikos nicht empfehlenswert ist, wird auf eine exakte Bestimmung des Risikos hier verzichtet.

5.6. Schätzung der Risiken bei der Verwendung von Kruskal's Streß

Da die Güte von NMDS-Darstellungen nicht nur hauptsächlich mit Hilfe von Kruskal's Streß beurteilt wird, sondern das Streßmaß auch entscheidend für die Bestimmung der optimalen Konfiguration ist, sollen hier Bayes-Risiken bestimmt werden, die sich aus Verlust- und Entscheidungsfunktionen ergeben, die auf Kruskal's Streß basieren.

Anders als bei der Verwendung des Permutationsabstandes können wir hier keine Verteilung von Stresses zwischen darstellbaren Rangfolgen bestimmen, da diese nicht definiert sind. Es ist also nicht möglich, eine Verteilung der darstellbaren Rangfolgen in der Gesamtmenge der Rangfolgen unter Verwendung des Streßes als Abstandsmaß zu bestimmen. Wir gehen deshalb von einem modifizierten Entscheidungsraum aus, dem auch ein anderes Erkenntnisinteresse entspricht:

Sei θ eine darstellbare Rangfolge. Dann ist eine Entscheidung $\varphi(x)$ mit $x \sim F_{\theta}(x)$ richtig, wenn $\varphi(x)$ die Entscheidung zugunsten irgendeiner darstellbaren Rangfolge ist, gleichgültig, ob $\theta = \varphi(x)$ oder $\theta \neq \varphi(x)$. D.h. die Hypothese H_0 lautet:
 $\theta \in \Theta_0 =$ Menge der darstellbaren Rangfolgen.

Diese Hypothesenbildung entspricht einem Gesichtspunkt, der die Darstellung per se erzeugen will, ohne die Fragestellung mit einzubeziehen, ob eine eventuell andere räumliche Darstellung adäquater wäre.

Als Gegenhypothesen wollen wir alle θ mit $\theta \notin \Theta_0$ zulassen, d.h. eine endliche Menge von Gegenhypothesen prüfen.

Die Entscheidungsfunktion zum Niveau α lautet hier:

$$\psi(x) = \begin{cases} d(x) \in D & \text{falls Streß}(x) \leq \alpha \\ x & \text{sonst} \end{cases}$$

Die Dichtefunktion der Zufallsvariablen X ist hier unterschiedlich für die Fälle $\theta \in \Theta_0$ und $\theta \notin \Theta_0$ zu bestimmen:

Fall 1: $\theta \in \Theta_0$. Dann geht in die Dichtefunktion $f_{\theta}^S(i) = f_{\theta}^S(x \in \{x\}_i)$; $i = \text{Streß}(x)$, die Verteilung der Streßwerte ein:

$$f_{\theta}^S(i) = \frac{\exp\left(-\frac{i^2}{2\sigma_2^2}\right) \cdot i^{p-1} \cdot (1-i)^{q-1}}{\int_0^1 \exp\left(-\frac{i^2}{2\sigma_2^2}\right) \cdot i^{p-1} \cdot (1-i)^{q-1} di}$$

wobei p und q die zur jeweiligen Punkte- und Dimensionszahl gehörigen Parameter der Beta-Verteilungen aus Tabelle 5.5 sind. $\sigma_2^2 = k^2 \cdot S_{\max}^2$ wie in Kapitel 5.5.

Fall 2:

$\theta \notin \Theta_0$. Dann ist θ nicht darstellbar und in die Dichte der Verteilungsfunktion um θ geht nicht die Verteilung der Streßwerte mit ein. Vielmehr ist es sinnvoll, eine Verteilung zu unterstellen, in die stattdessen die Verteilungen der Permutationen um θ mit eingeht. In diesem Fall ist $f_{\theta}(i)$ deshalb identisch mit (5.4).

Die Verlustfunktion bestimmt sich im 1. Fall aus der Größe des Streßes von x ,

$$\text{d.h. } L(\theta, \psi(x)) = \begin{cases} \text{Streß}(x) & \text{für } x \notin A_j \\ 0 & \text{falls } x \in A_j \end{cases}$$

Im 2. Fall wird wie in 5.5 der Permutationsabstand benutzt, da θ nicht darstellbar ist.

Wir erhalten daher folgende Risikofunktionen:

$$\hat{R}_1^s(\varphi, \theta) = \int_0^1 f_{\theta}^s(i) \cdot i \cdot di \quad ; \quad i = \text{Streß}$$

$$\hat{R}_2(\varphi, \theta) = P(x \in \bigcup_j A_j) \cdot \int_0^1 f_{\theta}(i) \cdot i \cdot di \quad ;$$

$$i = s(\varphi(x), \theta)$$

Die Wahrscheinlichkeit $\hat{H}(\theta \in D) = \hat{H}(D)$ wird mit Hilfe der Beta(p,q)-Verteilungen geschätzt. Damit kann das Bayes-Risiko bestimmt werden:

$$(5.9) \quad \hat{R}^s(\varphi, H) = \hat{H}(D) \hat{R}_1^s(\varphi, \theta) + (1 - \hat{H}(D)) \cdot \hat{R}_2(\varphi, \theta)$$

Tabelle 5.10 enthält Werte für $\hat{R}^s(\varphi, \theta)$ für verschiedene Punktezahlen n, Dimensionen p, Entscheidungskriterien α und Fehlerkoeffizienten k. Die Werte für α entsprechen hier Streßwerten, bei denen Entscheidungen für darstellbare Rangfolgen getroffen werden. Wie bei $\hat{R}(\varphi, H)$ kann eine Robustheit gegenüber Veränderungen von k festgestellt werden. Anders als bei $\hat{R}(\varphi, H)$ sind bei $\hat{R}^s(\varphi, H)$ größere Sprünge von niedrigen zu hohen Risikowerten festzustellen bei Vergrößerung von α oder n. Dies ist auf geringere Varianzen von $f_{\theta}^s(i)$ im Verhältnis zu $f_{\theta}(i)$ zurückzuführen. Das Zugehörige Programm ist eine (geringfügige) Modifikation von Programm 3 und daher hier nicht wiedergegeben.

Will man anhand von Tabelle 5.9 wiederum die notwendigen Mindestzahlen von Punkten für Darstellungen in 1 - 5 Dimensionen bei einem maximalen Streß von $\alpha \leq .15$ und einem Risiko unter 0.01 feststellen,

so ergibt sich:

Tabelle 5.11

Dim.	notwendige Mindestpunktezahl
1	7
2	11
3	20
4	25
5	35

Die erheblichen Unterschiede zu Tabelle 5.9 erklären sich aus den Verteilungen der Streßwerte, die geringere Erwartungswerte haben als die Permutationsabstände. Hierdurch wachsen die Wahrscheinlichkeiten für Fehler 1. Art, während^{sie} bei Fehlern 2. Art dieselben bleiben aufgrund der identischen Definition der Risikofunktion $R_2(\psi, \theta)$. Für eine befriedigendere Formulierung einer Risikofunktion müßte ein NMDS-Verfahren entwickelt werden, das die Minimierung des Permutationsabstandes anstelle des Streßes zum Ziel hat. Bei der Benutzung von MINISSA empfiehlt der Gebrauch von Tabellen 5.10 und 5.11, die konservativere Voraussetzungen haben als Tabellen 5.8 und 5.9.

Bei der Fehlerabschätzung sind die gleichen Punkte zu berücksichtigen, die bereits in Kapitel 5.5 erwähnt wurden. Nennenswerte Fehler treten nur für kleine Punktezahlen ($n \leq 5$) und kleine Punktezahlen im Verhältnis zu den Dimensionen ($n \leq p+2$) auf. Auf eine exakte Bestimmung der Risiken in diesen Fällen wird hier verzichtet, da eine Skalierung dort ohnehin nicht ratsam ist.

TABELLE 5.10

BAYES-RISIKEN UNTER EINBEZIEHUNG VON KRUSKALS STRESS GESCHAETZT

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ K=1.0000)

PKT	DIM	ALPHA							
		0.00	.01	.05	.10	.15	.20	.25	.30
4	1	.4244	.4554	.4554	.4518	.4455	.4363	.4237	.4100
5	1	.0003	.0007	.0456	.1674	.2903	.3809	.4369	.4673
	2	.4520	.4554	.4540	.4498	.4429	.4330	.4195	.4049
6	1	.0000	.0000	.0000	.0030	.0473	.1830	.3457	.4470
	2	.3568	.4414	.4579	.4583	.4540	.4466	.4362	.4247
	3	.4517	.4555	.4541	.4498	.4430	.4331	.4195	.4050
7	1	.0000	.0000	.0000	.0000	.0006	.0212	.1379	.3324
	2	.0000	.0000	.0945	.3823	.4824	.4958	.4969	.4969
	3	.4224	.4560	.4571	.4532	.4467	.4375	.4249	.4113
	4	.4541	.4552	.4538	.4495	.4427	.4327	.4191	.4045
8	1	.0000	.0000	.0000	.0000	.0000	.0006	.0240	.1725
	2	.0000	.0000	.0001	.1024	.4373	.4965	.4978	.4978
	3	.2729	.3928	.4613	.4681	.4652	.4594	.4516	.4426
	4	.4396	.4562	.4554	.4512	.4446	.4350	.4218	.4076
	5	.4537	.4553	.4539	.4496	.4427	.4328	.4192	.4046
9	1	.0000	.0000	.0000	.0000	.0000	.0000	.0013	.0508
	2	.0000	.0000	.0000	.0022	.2410	.4689	.4983	.4983
	3	.0000	.0000	.1503	.4810	.4982	.4983	.4983	.4983
	4	.3363	.4330	.4659	.4630	.4578	.4505	.4406	.4296
	5	.4447	.4560	.4549	.4507	.4439	.4342	.4209	.4066
10	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0047
	2	.0000	.0000	.0000	.0000	.0414	.4417	.4985	.4987
	3	.0000	.0000	.0010	.3989	.4985	.4987	.4987	.4987
	4	.1690	.2354	.4633	.4796	.4783	.4747	.4699	.4643
	5	.3490	.4388	.4649	.4616	.4562	.4487	.4384	.4270
11	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	2	.0000	.0000	.0000	.0000	.0008	.2814	.4979	.4989
	3	.0000	.0000	.0000	.1681	.4985	.4989	.4989	.4989
	4	.0000	.0000	.2186	.4979	.4989	.4989	.4989	.4989
	5	.1675	.2834	.4800	.4814	.4787	.4751	.4703	.4647
12	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0841	.4927	.4991
	3	.0000	.0000	.0000	.0148	.4944	.4991	.4991	.4991
	4	.0000	.0000	.0079	.4969	.4991	.4991	.4991	.4991
	5	.0244	.0496	.4517	.4959	.4962	.4956	.4950	.4941

FORTSETZUNG TABELLE 5.10

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K=1.0000$)

PKT	DIM	ALPHA							
		0.00	0.01	0.05	0.10	0.15	0.20	0.25	0.30
13	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0075	.4673	.4993
	3	.0000	.0000	.0000	.0001	.4573	.4993	.4993	.4993
	4	.0000	.0000	.0000	.4774	.4993	.4993	.4993	.4993
	5	.0000	.0000	.2940	.4992	.4993	.4993	.4993	.4993
14	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0001	.3784	.4994
	3	.0000	.0000	.0000	.0000	.3080	.4994	.4994	.4994
	4	.0000	.0000	.0000	.3621	.4994	.4994	.4994	.4994
	5	.0000	.0000	.0349	.4994	.4994	.4994	.4994	.4994
15	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.2000	.4994
	3	.0000	.0000	.0000	.0000	.0927	.4995	.4995	.4995
	4	.0000	.0000	.0000	.1249	.4995	.4995	.4995	.4995
	5	.0000	.0000	.0002	.4993	.4995	.4995	.4995	.4995
20	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.2615
	3	.0000	.0000	.0000	.0000	.0000	.1625	.4997	.4997
	4	.0000	.0000	.0000	.0000	.2664	.4997	.4997	.4997
	5	.0000	.0000	.0000	.0037	.4997	.4997	.4997	.4997
25	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0018
	3	.0000	.0000	.0000	.0000	.0000	.0000	.4996	.4998
	4	.0000	.0000	.0000	.0000	.0001	.4998	.4998	.4998
	5	.0000	.0000	.0000	.0000	.4958	.4998	.4998	.4993
30	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.4352	.4999
	4	.0000	.0000	.0000	.0000	.0000	.4934	.4999	.4999
	5	.0000	.0000	.0000	.0000	.1573	.4999	.4999	.4999
35	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.1219	.4999
	4	.0000	.0000	.0000	.0000	.0000	.3216	.4999	.4999
	5	.0000	.0000	.0000	.0000	.0000	.4999	.4999	.4999
40	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.2145
	4	.0000	.0000	.0000	.0000	.0000	.1189	.4790	.4998
	5	.0000	.0000	.0000	.0000	.0000	.3599	.4998	.4999

FORTSETZUNG TABELLE 5.10

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K = .5000$)

PKT	DIM	ALPHA							
		.000	.01	.05	.10	.15	.20	.25	.30
4	1	.3319	.3605	.3599	.3545	.3457	.3331	.3167	.2992
5	1	.0003	.0006	.0436	.1601	.2776	.3642	.4178	.4469
	2	.3534	.3568	.3546	.3488	.3392	.3257	.3080	.2894
6	1	.0000	.0000	.0000	.0029	.0461	.1781	.3364	.4351
	2	.2793	.3616	.3772	.3761	.3698	.3593	.3458	.3309
	3	.3532	.3569	.3549	.3469	.3393	.3259	.3082	.2896
7	1	.0000	.0000	.0000	.0000	.0006	.0208	.1354	.3263
	2	.0000	.0000	.0927	.3753	.4736	.4868	.4878	.4879
	3	.3304	.3634	.3639	.3583	.3493	.3368	.3203	.3029
	4	.3550	.3562	.3542	.3482	.3385	.3250	.3073	.2886
8	1	.0000	.0000	.0000	.0000	.0000	.0005	.0237	.1703
	2	.0000	.0000	.0001	.1010	.4315	.4900	.4912	.4912
	3	.2138	.3320	.3493	.4049	.4003	.3924	.3820	.3706
	4	.3438	.3602	.3586	.3529	.3436	.3305	.3134	.2952
	5	.3547	.3563	.3543	.3483	.3387	.3252	.3074	.2887
9	1	.0000	.0000	.0000	.0000	.0000	.0000	.0013	.0502
	2	.0000	.0000	.0000	.0321	.2386	.4840	.4933	.4933
	3	.0000	.0000	.1488	.4762	.4932	.4933	.4933	.4933
	4	.2632	.3590	.3911	.3869	.3796	.3696	.3567	.3427
	5	.3477	.3590	.3573	.3513	.3419	.3280	.3113	.2929
10	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0047
	2	.0000	.0000	.0000	.0000	.2411	.4382	.4946	.4947
	3	.0000	.0000	.0010	.3958	.4946	.4947	.4947	.4947
	4	.1325	.2479	.4242	.4399	.4373	.4324	.4261	.4189
	5	.2732	.3622	.3877	.3831	.3755	.3652	.3518	.3372
11	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	2	.0000	.0000	.0000	.0000	.0008	.2796	.4947	.4957
	3	.0000	.0000	.0000	.1670	.4953	.4957	.4957	.4957
	4	.0000	.0000	.2172	.4948	.4957	.4957	.4957	.4957
	5	.1312	.2464	.4415	.4422	.4385	.4336	.4273	.4202
12	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0836	.4901	.4965
	3	.0000	.0000	.0000	.0148	.4918	.4965	.4965	.4965
	4	.0000	.0000	.0078	.4943	.4965	.4965	.4965	.4965
	5	.0191	.0442	.4442	.4880	.4881	.4874	.4865	.4855

FORTSETZUNG TABELLE 5.10

(KOLLEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K=0.5000$)

PKT	DIM	ALPHA							
		0.00	.01	.05	.10	.15	.20	.25	.30
13	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0074	.4652	.4970
	3	.0000	.0000	.0000	.0001	.4553	.4970	.4970	.4970
	4	.0000	.0000	.0000	.4753	.4970	.4970	.4970	.4970
	5	.0000	.0000	.2927	.4970	.4970	.4970	.4970	.4970
14	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0001	.3770	.4975
	3	.0000	.0000	.0000	.0000	.3068	.4975	.4975	.4975
	4	.0000	.0000	.0000	.3607	.4975	.4975	.4975	.4975
	5	.0000	.0000	.0348	.4975	.4975	.4975	.4975	.4975
15	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.2053	.4978
	3	.0000	.0000	.0000	.0000	.0924	.4978	.4978	.4978
	4	.0000	.0000	.0000	.1245	.4978	.4978	.4978	.4978
	5	.0000	.0000	.0002	.4977	.4978	.4978	.4978	.4978
20	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.2611
	3	.0000	.0000	.0000	.0000	.0000	.1622	.4988	.4988
	4	.0000	.0000	.0000	.0000	.2660	.4988	.4988	.4988
	5	.0000	.0000	.0000	.0037	.4988	.4988	.4988	.4988
25	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0018
	3	.0000	.0000	.0000	.0000	.0000	.0000	.4990	.4993
	4	.0000	.0000	.0000	.0000	.0001	.4993	.4993	.4993
	5	.0000	.0000	.0000	.0000	.4952	.4993	.4993	.4993
30	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.4349	.4995
	4	.0000	.0000	.0000	.0000	.0000	.4930	.4995	.4995
	5	.0000	.0000	.0000	.0000	.1572	.4995	.4995	.4995
35	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.1219	.4996
	4	.0000	.0000	.0000	.0000	.0000	.3216	.4996	.4996
	5	.0000	.0000	.0000	.0000	.0008	.4996	.4996	.4996
40	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.1953
	4	.0000	.0000	.0000	.0000	.0000	.2139	.4788	.4995
	5	.0000	.0000	.0000	.0000	.0000	.3067	.4995	.4997

FORTSETZUNG TABELLE 5.10

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K=1.0000$)

PKT	DIM	ALPHA							
		0.00	.01	.05	.10	.15	.20	.25	.30
4	1	.0693	.0781	.0706	.0514	.0318	.0185	.0120	.0098
5	1	.0001	.0002	.0182	.0672	.1166	.1530	.1755	.1877
	2	.0734	.0749	.0665	.0461	.0252	.0111	.0042	.0018
6	1	.0000	.0000	.0000	.0016	.0248	.0959	.1812	.2343
	2	.0589	.1032	.1056	.0912	.0752	.0641	.0588	.0569
	3	.0734	.0754	.0670	.0465	.0257	.0116	.0047	.0023
7	1	.0000	.0000	.0000	.0000	.0004	.0132	.0856	.2062
	2	.0000	.0000	.0586	.2372	.2993	.3076	.3083	.3083
	3	.0690	.0897	.0833	.0641	.0446	.0313	.0249	.0226
	4	.0737	.0744	.0659	.0454	.0244	.0102	.0033	.0009
8	1	.0000	.0000	.0000	.0000	.0000	.0004	.0167	.1197
	2	.0000	.0000	.0000	.0710	.3033	.3444	.3452	.3452
	3	.0455	.1285	.1713	.1652	.1533	.1448	.1407	.1392
	4	.0716	.0829	.0751	.0552	.0348	.0211	.0144	.0120
	5	.0737	.0746	.0662	.0456	.0247	.0105	.0036	.0012
9	1	.0000	.0000	.0000	.0000	.0000	.0000	.0010	.0380
	2	.0000	.0000	.0000	.0016	.1805	.3662	.3733	.3733
	3	.0000	.0000	.1126	.3603	.3732	.3733	.3733	.3733
	4	.0555	.1278	.1468	.1316	.1159	.1053	.1001	.0983
	5	.0723	.0806	.0725	.0523	.0318	.0179	.0111	.0088
10	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0036
	2	.0000	.0000	.0000	.0000	.0328	.3497	.3947	.3948
	3	.0000	.0000	.0008	.3158	.3947	.3948	.3948	.3948
	4	.0284	.1204	.2584	.2648	.2578	.2525	.2499	.2490
	5	.0575	.1283	.1433	.1273	.1110	.1001	.0947	.0928
11	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	2	.0000	.0000	.0000	.0000	.0007	.2321	.4107	.4115
	3	.0000	.0000	.0000	.1387	.4111	.4115	.4115	.4115
	4	.0000	.0000	.1803	.4107	.4115	.4115	.4115	.4115
	5	.0281	.1236	.2828	.2775	.2696	.2643	.2618	.2609
12	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0715	.4192	.4247
	3	.0000	.0000	.0000	.0126	.4206	.4247	.4247	.4247
	4	.0000	.0000	.0067	.4228	.4247	.4247	.4247	.4247
	5	.0042	.0256	.3673	.4039	.4033	.4025	.4022	.4020

FORTSETZUNG TABELLE 5.10

(KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ $K=1.0000$)

PKT	DIM	ALPHA							
		0.00	.01	.05	.10	.15	.20	.25	.30
13	1	0.00000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0065	.4073	.4352
	3	.0000	.0000	.0000	.0000	.3486	.4352	.4352	.4352
	4	.0000	.0000	.0000	.4161	.4352	.4352	.4352	.4352
	5	.0000	.0000	.2563	.4352	.4352	.4352	.4352	.4352
14	1	0.00000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	0.00000	.0000	.0000	.0000	.0000	.0001	.3363	.4437
	3	.0000	.0000	.0000	.0000	.2736	.4437	.4437	.4437
	4	.0000	.0000	.0000	.3217	.4437	.4437	.4437	.4437
	5	.0000	.0000	.0310	.4437	.4437	.4437	.4437	.4437
15	1	0.00000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	0.00000	.0000	.0000	.0000	.0000	.0000	.1859	.4507
	3	.0000	.0000	.0000	.0000	.0836	.4507	.4507	.4507
	4	.0000	.0000	.0000	.1127	.4507	.4507	.4507	.4507
	5	.0000	.0000	.0002	.4506	.4507	.4507	.4507	.4507
20	1	0.00000	.00000	.00000	.0000	.0000	.0000	.0000	.0000
	2	0.00000	.00000	.0000	.0000	.0000	.0000	.0000	.2470
	3	0.00000	.0000	.0000	.0000	.0000	.1534	.4719	.4719
	4	0.00000	.0000	.0000	.0000	.2516	.4719	.4719	.4719
	5	.0000	.0000	.0000	.0035	.4719	.4719	.4719	.4719
25	1	0.00000	.00000	.00000	.00000	.0000	.0000	.0000	.0000
	2	0.00000	.00000	.0000	.0000	.0000	.0000	.0000	.0017
	3	0.00000	.00000	.0000	.0000	.0000	.0000	.4817	.4819
	4	0.00000	.0000	.0000	.0000	.0001	.4819	.4819	.4819
	5	0.00000	.0000	.0000	.0000	.4781	.4819	.4819	.4819
30	1	0.00000	.00000	.00000	.00000	.00000	.00000	.0000	.0000
	2	0.00000	.00000	.00000	.0000	.0000	.0000	.0000	.0000
	3	0.00000	.00000	.0000	.0000	.0000	.0000	.4244	.4875
	4	0.00000	.0000	.0000	.0000	.0000	.4812	.4875	.4875
	5	0.00000	.0000	.0000	.0000	.1534	.4875	.4875	.4875
35	1	0.00000	.00000	.00000	.00000	.00000	.00000	.0000	.0000
	2	0.00000	.00000	.00000	.00000	.0000	.0000	.0000	.0000
	3	0.00000	.00000	.0000	.0000	.0000	.0000	.1197	.4906
	4	0.00000	.00000	.0000	.0000	.0000	.3158	.4906	.4906
	5	0.00000	.0000	.0000	.0000	.0008	.4906	.4906	.4906
40	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.3953
	4	.0000	.0000	.0000	.0000	.0000	.2156	.4717	.4922
	5	.0000	.0000	.0000	.0000	.0000	.3034	.4922	.4923

6. Kapitel

Weitere Ergebnisse der Monte-Carlo Studie

Zum Abschluß sollen noch einige weitere Statistiken zu den empirisch erhobenen Maßen dargestellt werden. Insbesondere interessieren die Zusammenhänge zwischen Kruskal's Streß und dem Permutationsabstand, die Aufschlüsse über den Grad, in dem das Verfahren das Isotoniekriterium einhält, geben können.

Zunächst sei aber die Abhängigkeit der Maße von der Anzahl der Punkte einer Konfiguration sowie der Anzahl der Dimensionen des Konfigurationsraumes untersucht.

Graphiken 1, 2 und 3 zeigen die durchschnittlichen Werte für Kruskal's Streß, den Alienationskoeffizienten und für das normierte Permutationsmaß, die jeweils mit 1000 multipliziert wurden. Die gemittelten Maße sind nach der Anzahl der Punkte abgetragen. Es zeigen sich für alle drei Maße deutliche funktionale Zusammenhänge zwischen den Mittelwerten und der Punktezahl. Ferner können jeweils fünf Kurven identifiziert werden, die zu den Dimensionen 1-5 gehören.

Graphiken 4-6 zeigen die Zusammenhänge zwischen den gemittelten Maßen und den fünf Dimensionen. Auch hier könnte man wiederum siebzehn nahezu parallel verlaufende Kurven ausmachen, die zu den verschiedenen Punktezahlen gehören.

Es bietet sich nun an, Regressionen der einzelnen Maße auf die Punktezahlen und Dimensionen zu berechnen. Eine Regression von den gemittelten Kruskal's Streßen auf die vier Variablen n , p , $\ln(n)$ und $\sqrt{\ln(p)}$, wie sie Spence (42) vorschlug, erbringt gute Ergebnisse.

Die Regressionsgleichung heie

$$(6.1) \quad \text{KSTR} = a_0 + a_1 n + a_2 p + a_3 \ln p + a_4 \sqrt{\ln(n)}$$

Bei einem R^2 von .9856 ergeben sich die Koeffizienten

$$a_0 = -945.98984$$

$$a_1 = -6.17206$$

$$a_2 = 38.42207$$

$$a_3 = 869.20885$$

$$a_4 = -313.95832$$

Das gleiche fr die gemittelten Guttman's Alienationskoeffizienten (GSTR) ergibt bei $R^2 = .9810$

$$a_0 = -1197.64184$$

$$a_1 = -8.74011$$

$$a_2 = 47.18885$$

$$a_3 = 1144.41248$$

$$a_4 = -412.29413$$

und die gleiche Regressionsgleichung fr die mittleren Permutationsabstnde ergibt bei $R^2 = .9410$

$$a_0 = -530.72166$$

$$a_1 = -1.39414$$

$$a_2 = -0.67355$$

$$a_3 = 546.16242$$

$$a_4 = -111.48544$$

Im Unterschied zu der hier vorgestellten Regression von KSTR steht eine Untersuchung von Spence (42). Dort ergab die Regression

$$\text{Kruskal's Stre} = a_0 + a_1 D + a_2 N + a_3 \ln D + a_4 \sqrt{\ln N}$$

mit D = Dimensionen und N = Punktezahl $N = (12, 18, 26, 36, 48)$ aus Stichproben von jeweils nur 15 Elementen unter Anwendung der MDS-Routine TORSCA die Koeffizienten $a_0 = -524.25$, $a_1 = 33.8$, $a_2 = -2.54$, $a_3 = -307.26$, $a_4 = 588.35$, $R^2 = 0.9963$.

Spence erhielt allerdings auch wesentlich hhere mittlere Strewerte aus seiner Stichprobe, als hier berechnet wurden (vgl. Tabellen 6.1 und 5.4). Die Unterschiede sind sicherlich weniger auf die geringeren Stichprobenumfnge als auf unterschiedliche Approximationsgenauig-

keiten der Verfahren TORSCA und MINISSA (SV 3) zurückzuführen.

Tabelle 6.1 : mittlere Streße nach Spence (42)

Stress Values from the Approximation and Errors of the Approximation					
Points	Dimensionality				
	1	2	3	4	5
12	407 (-001)	227 (-005)	187 (-006)	082 (+000)	047 (+004)
18	464 (+005)	285 (+006)	194 (+003)	140 (+004)	105 (+002)
26	505 (+002)	328 (+001)	236 (-002)	181 (-003)	146 (-002)
36	532 (-005)	353 (-001)	262 (-002)	207 (-003)	173 (-005)
48	545 (-002)	366 (+004)	275 (+004)	221 (+001)	186 (+001)

Note.—Unit of measurement is 0.001. Entries in brackets are the differences between the original Monte Carlo means and the approximation values.

Nimmt man in unserer Untersuchung ein etwas geringeres R^2 in Kauf, so läßt sich die Anzahl der unabhängigen Variablen auf zwei reduzieren. Dabei werden die Maße exponentiiert. Die Regressionsgleichung für die mittleren Kruskal's Streße heißt dann

$$(6.2) \quad \exp(KSTR) = a_0 + a_1 \ln(n) + a_2 \ln(p)$$

$$\text{bzw. } KSTR = \ln(\ln(e^{a_0} \cdot n^{a_1} \cdot p^{a_2}))$$

Mit $R^2 = .9360$ erhält man

$$a_0 = .92824$$

$$a_1 = .20099$$

$$a_2 = -.27386$$

Für die mittleren Permutationsabstände ergibt die gleiche Regressionsgleichung ein R^2 von .9536 und die Koeffizienten

$$a_0 = .90636$$

$$a_1 = .18061$$

$$a_2 = -.13756$$

Wir wollen nun Korrelationen zwischen den Maßen untersuchen (Tabellen 6.2 und 6.3).

Während Kruskal's Streß und Guttman's Alienationskoeffizienten stark korreliert sind (Tabelle 6.2) für alle Punktezahlen und Dimensionen, ist der Zusammenhang zwischen Kruskal's Streß und dem normierten Permutationsabstand deutlich schwächer (Tabelle 6.3). Graphiken 7 - 11 zeigen exemplarisch, daß ein nicht-linearer Zusammenhang auch nicht sichtbar wird.

Allerdings wird für geringe p die Korrelation stärker. Es ist daher festzustellen, daß das Streßmaß im allgemeinen nicht geeignet ist, die Isotonie zu optimieren. Das heißt, daß bei der Anwendung des Streßmaßes nicht erreicht wird, daß die Anzahl der Verletzungen der Isotonie minimiert wird. Vielmehr wird auch z.B. anhand der Graphiken 6.7-6.11 deutlich, daß geringe Streßwerte bei hohen Permutationsabständen möglich sind.

Aus Tabelle 5.4 ist ersichtlich, daß die mittleren Permutationsabstände generell höher sind als die Streßwerte, falls die Anzahl der Dimensionen größer ist als eins (bei Dimension = 1 kehrt sich das Verhältnis um: Hier sind die Verluste, in "Distanzen" gemessen größer, als wenn man sie in Permutationsschritten ausdrückt.)

Als letztes seien Hinweise auf lokale Minima zu vermerken. Lokale Minima sind dann anzunehmen, wenn die Streßwerte einer Konfiguration zunehmen, obwohl die Dimension des Raumes, in dem die Punkte eingebettet sind, erhöht wird. Dieses ereignete sich in dieser Untersuchung jedoch nur in den Fällen, in denen kleine Punktezahlen in Räumen mit relativ vielen Dimensionen skaliert werden sollten. Tabelle 6.4 gibt die Anzahl der Konfigurationen von den jeweils 1000 untersuchten wieder, die in der entsprechenden Dimension höhere Streßwerte aufwiesen als in Räumen geringerer Dimension.

Tabelle 6.2

Pearson-Korrelationen zwischen Kruskal's Streß und
Guttman's Alienationskoeffizienten

Pkt	Dim: 1	2	3	4	5
4	.798	-	-	-	-
5	.759	.999	-	-	-
6	.891	.998	.999	-	-
7	.872	.987	.998	.993	-
8	.876	.744	.997	.998	.999
9	.883	.893	.982	.998	.998
10	.851	.851	.941	.997	.998
11	.864	.885	.890	.976	.997
12	.863	.794	.952	.934	.995
13	.844	.785	.819	.893	.969
14	.831	.771	.796	.857	.932
15	.849	.745	.775	.814	.886
20	.858	.916	.932	.931	.941
25	.799	.689	.714	.762	.823
30	.766	.695	.648	.717	.745
35	.819	.655	.648	.707	.709
40	.845	.839	.720	.766	.790

Tabelle 6.3

Pearson-Korrelationen zwischen Kruskal's Streß
und dem Permutationsabstand

Pkt	Dim: 1	2	3	4	5
4	.638	-	-	-	-
5	.853	.935	-	-	-
6	.714	.415	.496	-	-
7	.678	.658	.321	.896	-
8	.687	.680	.584	.548	.345
9	.613	.602	.534	.339	.486
10	.798	.634	.519	.510	.335
11	.817	.675	.584	.370	-.005
12	.615	.636	.586	.508	-.076
13	.590	.685	.607	.535	.092
14	.605	.597	.638	.533	.378
15	.597	.677	.369	.535	.485
20	.517	.707	.615	.491	.139
25	.782	.629	.586	.415	.480
30	.734	.539	.293	.241	.126
35	.771	.569	.185	.157	.158
40	.821	.619	.397	.361	.352

Die entsprechenden Streßwerte wurden vor der Untersuchung aus dem Datensatz entfernt.

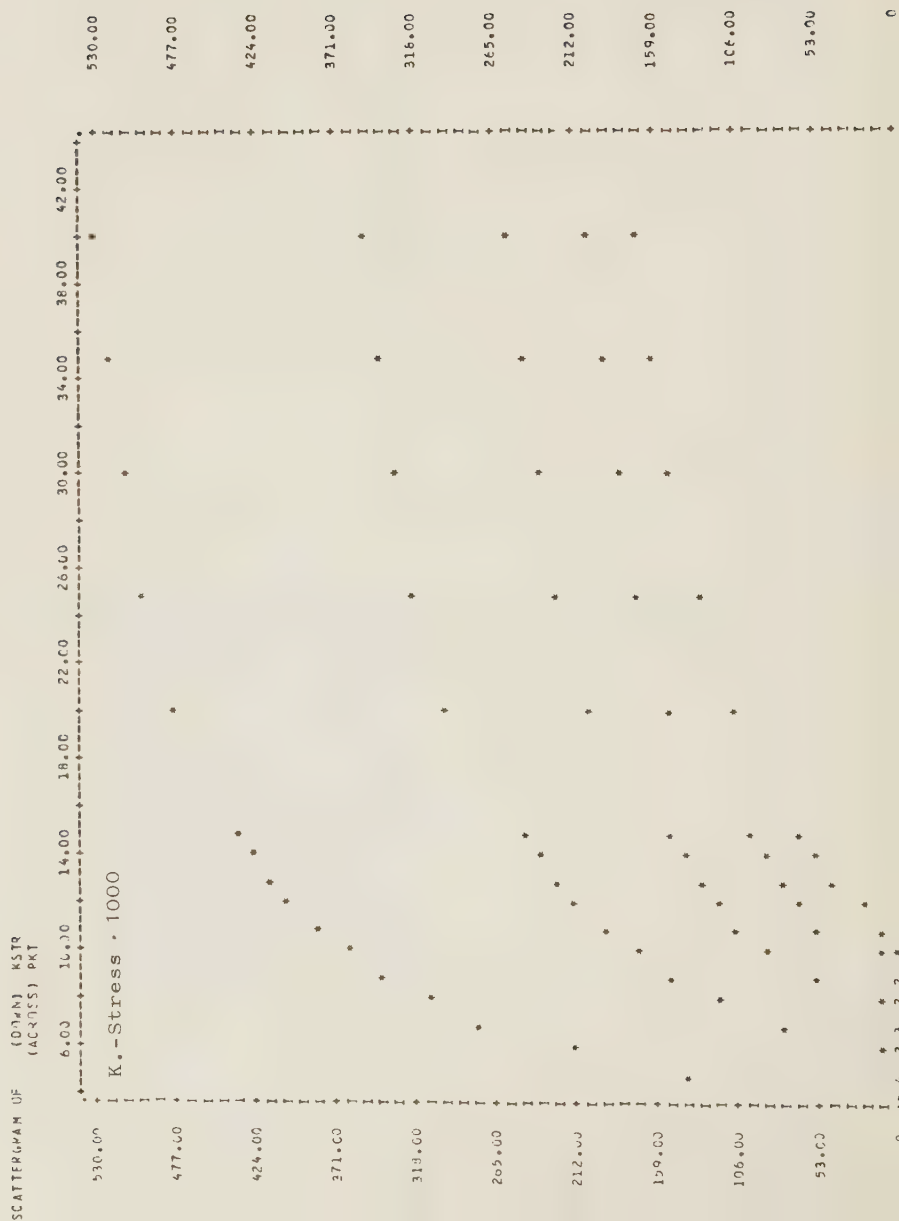
Tabelle 6.4

Anzahl der Konfigurationen, bei denen Kruskal's Streß für Dimension p größer war als bei der Skalierung im Raum mit $p-1$ Dimensionen (Lokale Minima).

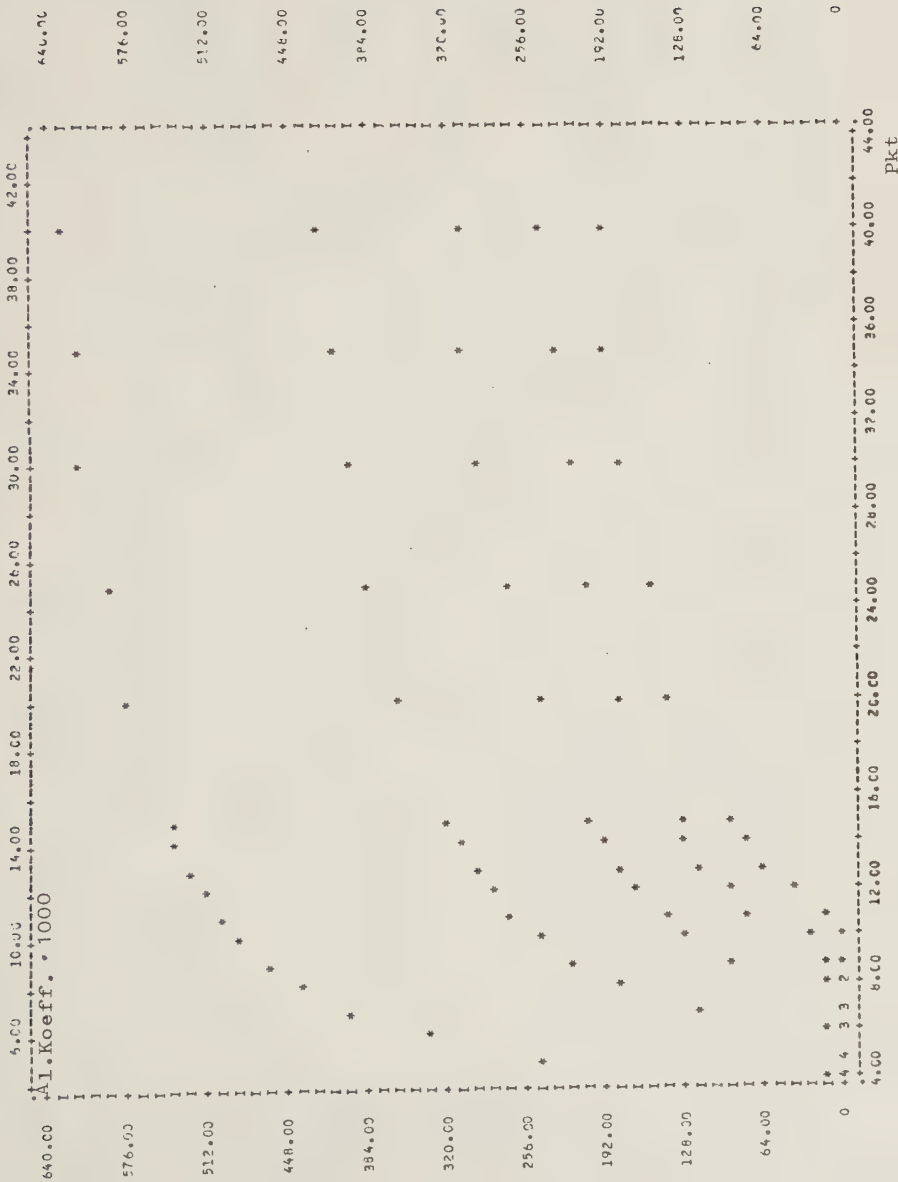
PKT	DIM			
	2	3	4	5
4	332	332	332	332
5	25	4	4	4
6	2	85	15	12
7	0	26	94	43
8	0	1	278	94
9	0	0	29	160
10	0	0	0	416
11	0	0	0	25
12	0	0	0	0
⋮	⋮	⋮	⋮	⋮
40	0	0	0	0

Für die Berechnungen und Graphiken wurden neben eigenen FORTRAN- Programmen die Programmbibliotheken SPSS (43) und BMDP (44) benutzt.

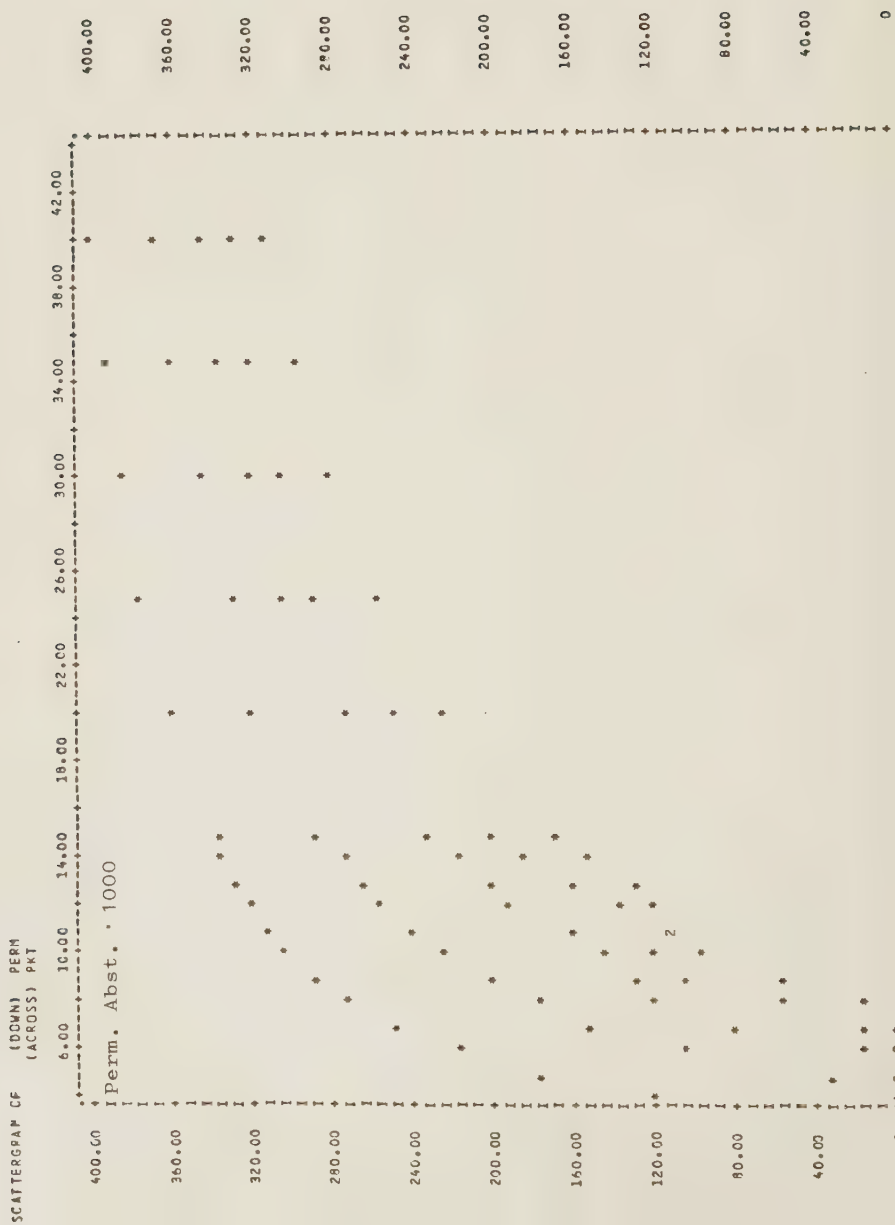
Graphik 1



Graphik 2

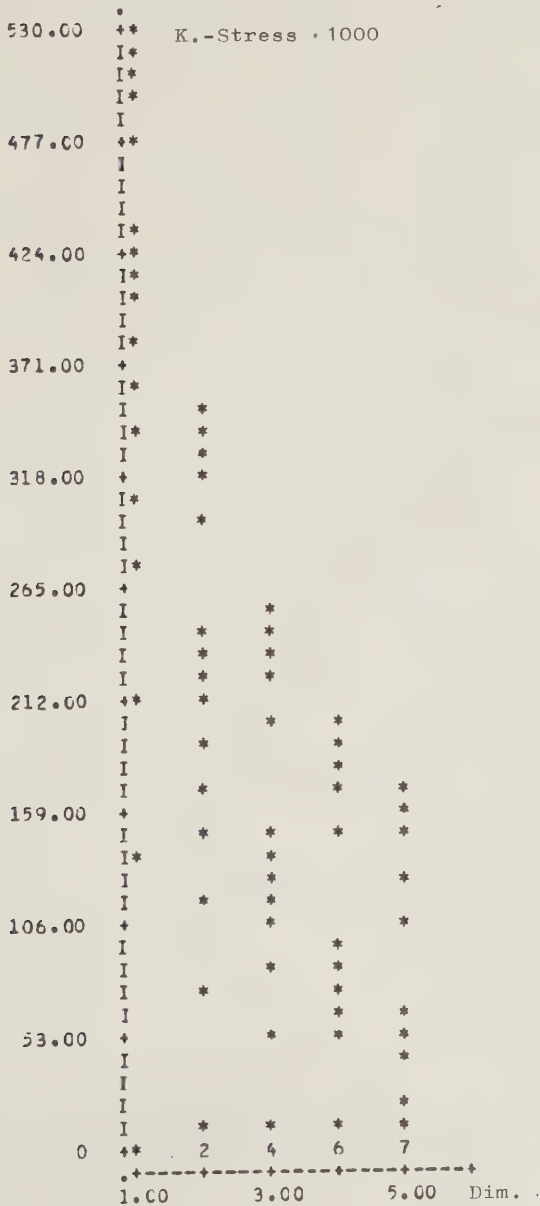


Graphik 3



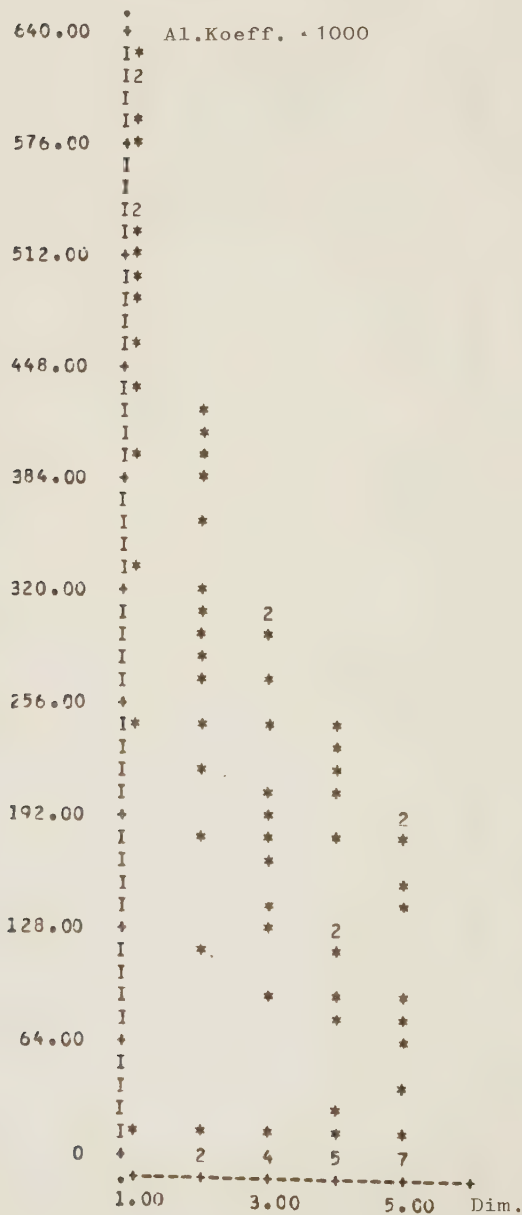
Graphik 4

SCATTERGRAM OF (DOWN) KSTR
(ACROSS) DIM



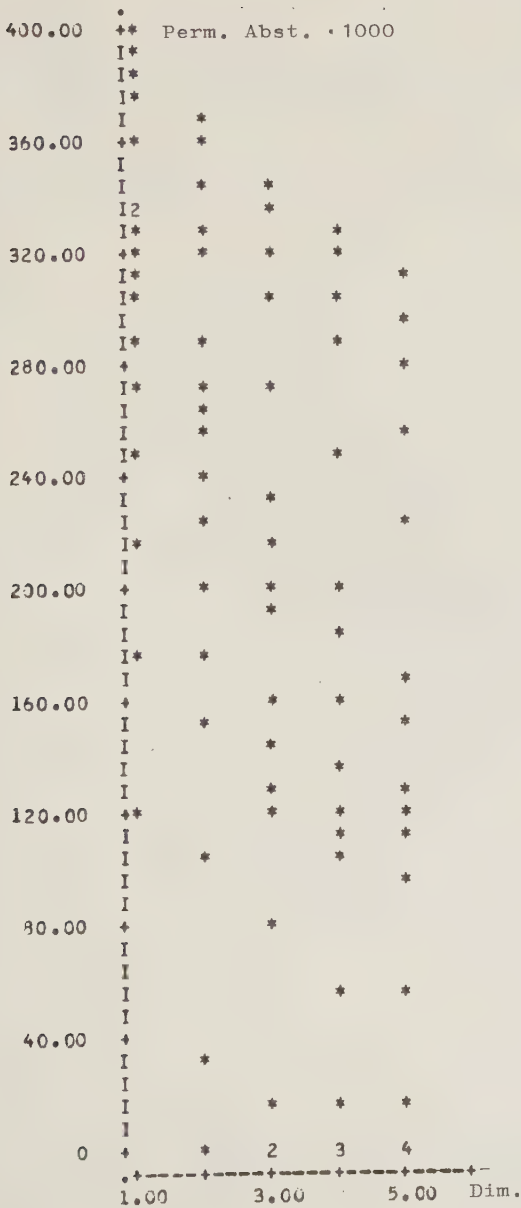
Graphik 5

SCATTERGRAM CF (DOWN) ESTR
(ACROSS) DIM



Graphik 6

SCATTERGRAM OF (DOWN) PERM
(ACROSS) DIM



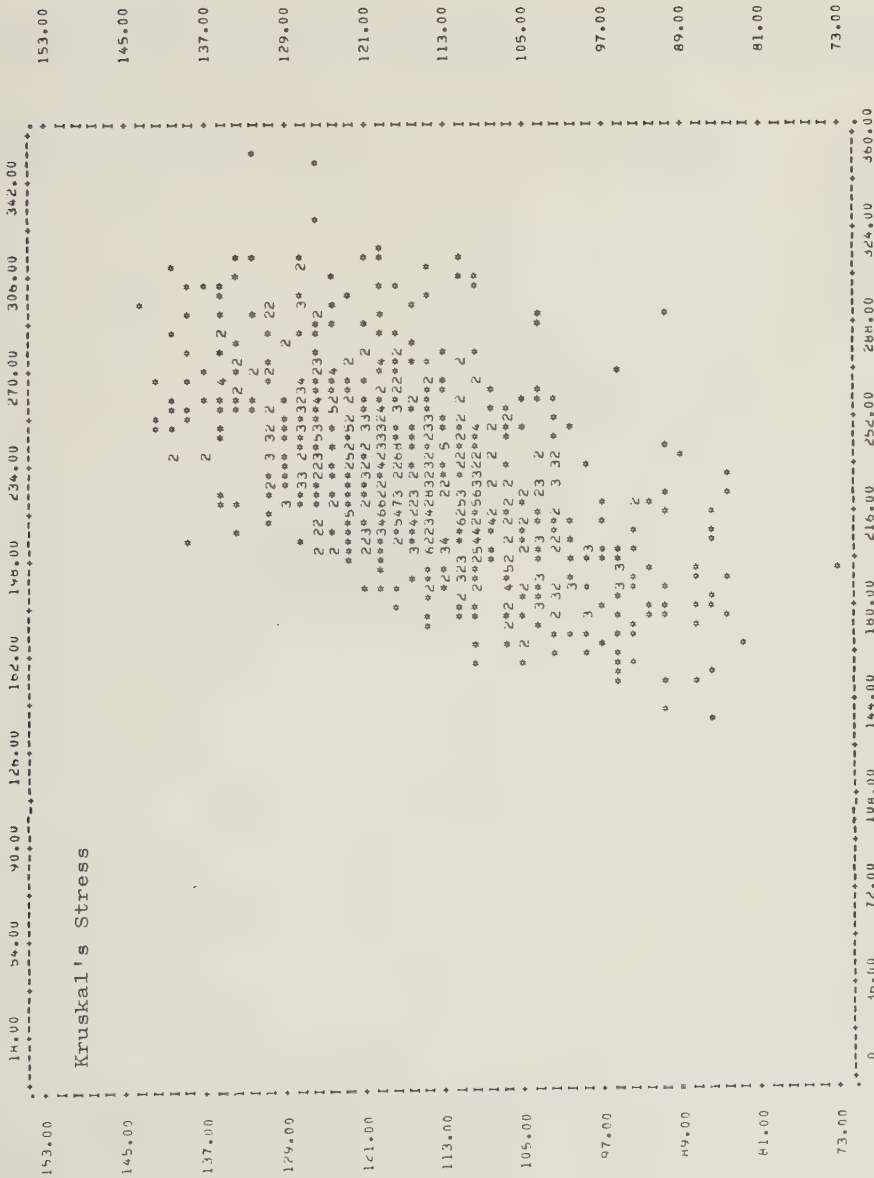
Graphik 7: Punktezahl=8
Dimensionen=1



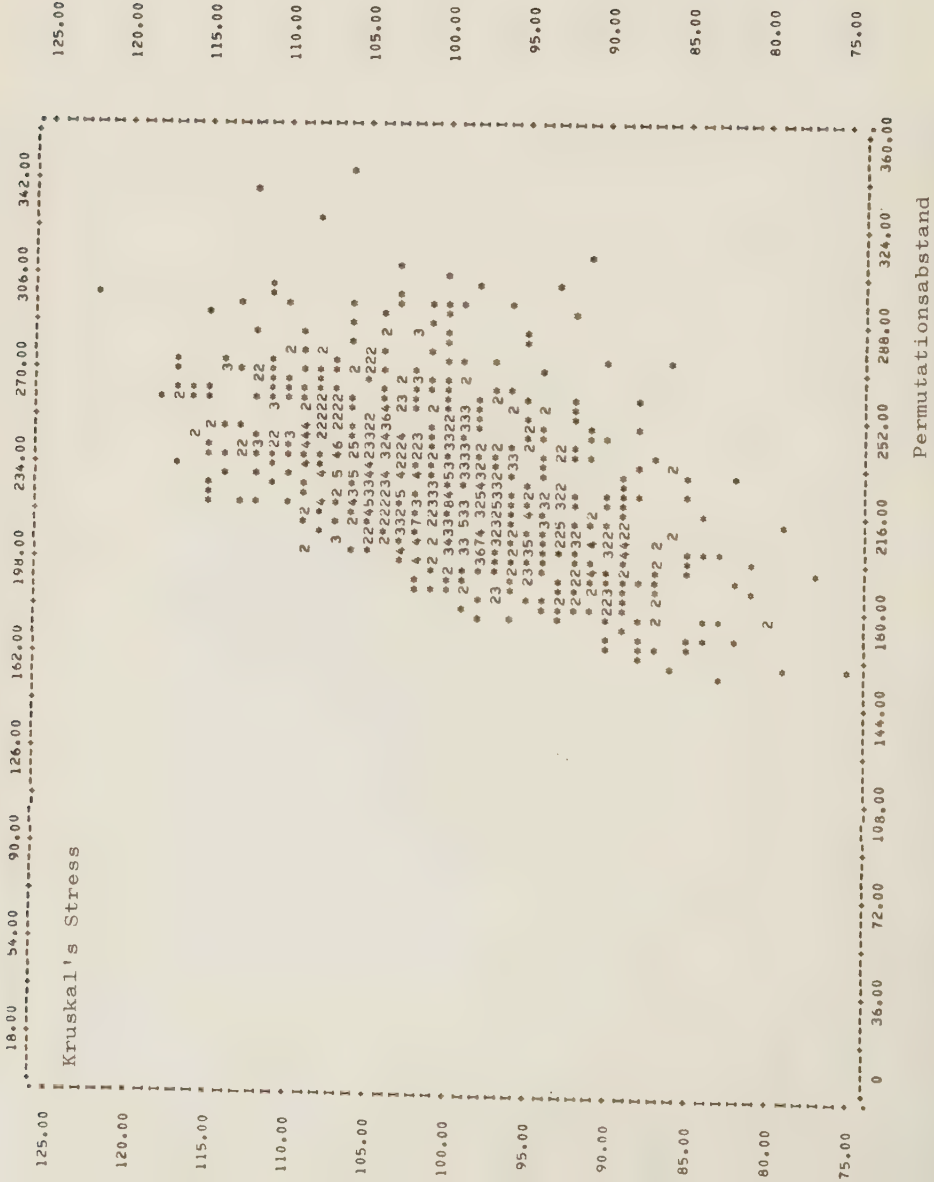
Graphik 8: Punktezahl = 12
Dimensionen = 3

SUNFILE P12
SCATTERGRAM OF
(U04N) KSTKJ
(ACHUSS) PETHJ

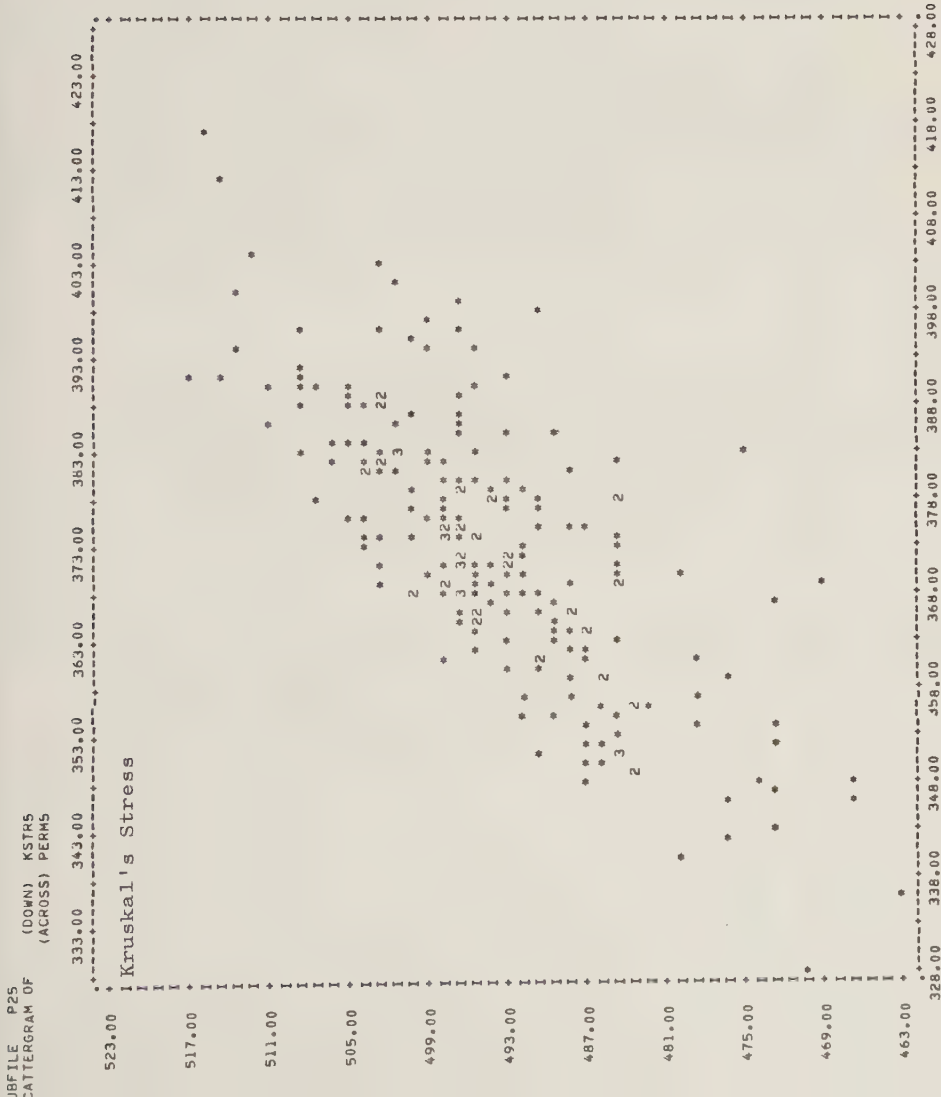
Kruskal's Stress



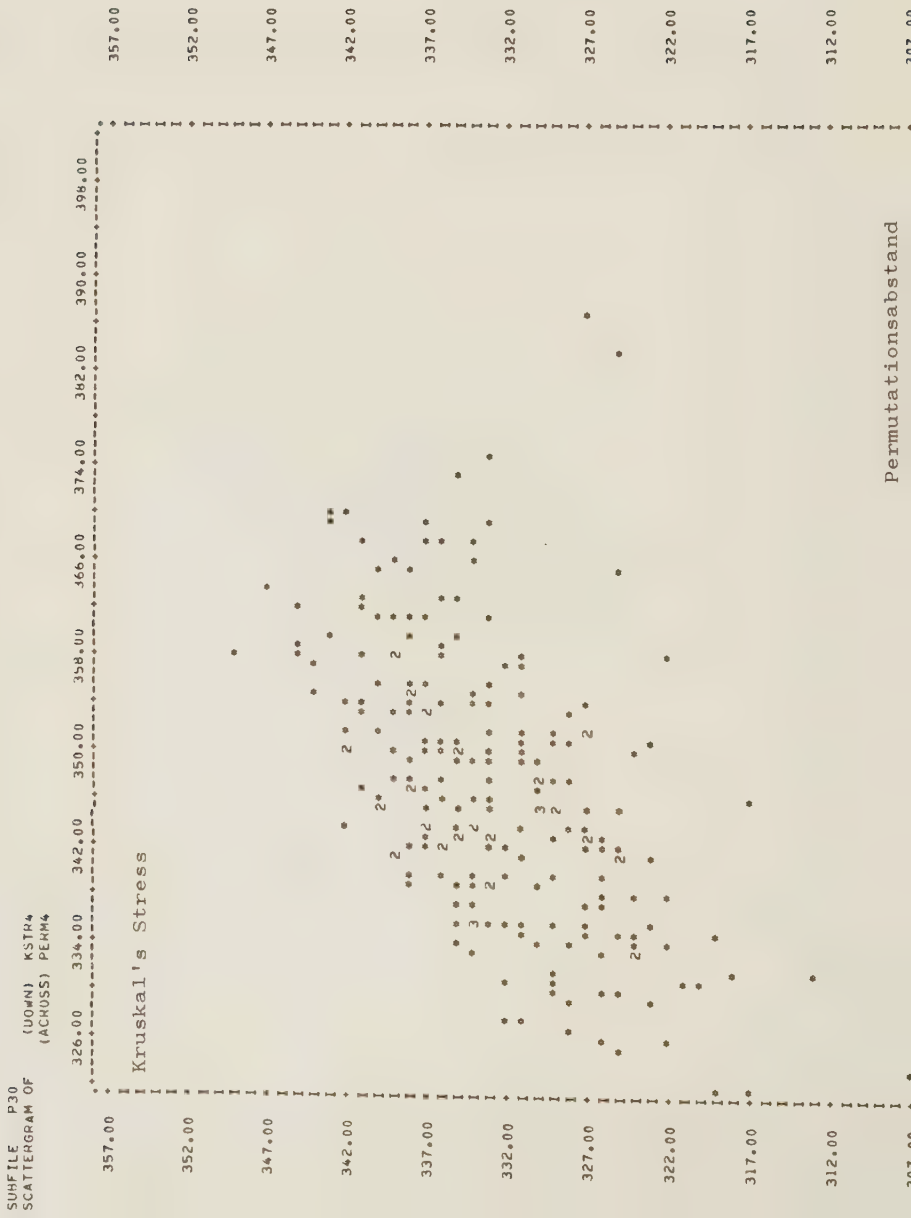
Graphik 9: Punktezahl = 15
Dimensionen = 4



Graphik 10: Punktezahl = 25
Dimensionen = 1



Graphik 11 : Punktezahl = 30
Dimensionen = 4



Anhang: Programm 1

```

*****
PROGRAMM ZUR BERECHNUNG DER VERTEILUNG VON ORDINALEN ABSTÄNDEN
ZWISCHEN PERMUTATIONEN UND ZUR SCHÄTZUNG DER ORDINALEN ABSTÄNDE
ZWISCHEN DARSTELLBAREN PERMUTATIONEN IN ABHÄNGIGKEIT VON
DIMENSION DES KONFIGURATIONENRAUMES UND DER PUNKTEZAHL.

```

```

MAXIMALE PUNKTEZAHL IST 40.
MAXIMALZAHL DER DIMENSIONEN IST 5.

```

EINGABE:

```

IM FORMAT 5I4 WERDEN AUF EINER KARTE ANGEGEBEN:

```

```

PMIN, PMAX, DMIN, DMAX, METRIK. BEDEUTUNG DER PARAMETER:

```

```

PMIN, DIE KLEINSTE ANZAHL VON PUNKTEN, DIE BERÜCKSICHTIGT WERDEN
SOLL 4< PMIN <40.

```

```

PMAX, DIE GRÖSSTE ANZAHL VON PUNKTEN; 4<PMAX<40.

```

```

DMIN, DIE KLEINSTE ANZAHL VON DIMENSIONEN; 0<DMIN<5.

```

```

DMAX, DIE GRÖSSTE ANZAHL VON DIMENSIONEN; 0<DMAX<5.

```

```

METRIK, DER KUEFFIZIENT DER MINKOWSKIMETRIK, 0<METRIK<99.

```

```

VORGABEWERTE SIND: 5 15 1 4 2

```

DAS PROGRAMM VOLLZIEHT FOLGENDE SCHRITTE:

1. GENERATION VON 1000 PAAREN VON ZUFALLSKONFIGURATIONEN MIT PMIN BIS PMAX PUNKTEN, DMIN BIS DMAX DIMENSIONEN.
2. BERECHNUNG DER DISTANZEN UNTER VERWENDUNG DES METRIKKOEFFIZIENTEN.
3. FÜR JEDES DER KONFIGURATIONENPAARE: BERECHNUNG DES ORDINALEN ABSTANDES (PERMUTATIONSABSTAND S)
4. BERECHNUNG DER VERTEILUNGEN DER ORDINALEN ABSTÄNDE

```

*****
PROGRAM VERT(INPUT,OUTPUT,TAPE1,TAPE2,TAPE3)
IMPLICIT INTEGER(A-Z)
DIMENSION R(5000),DIS1(800),DIS2(800),R1(800),R2(800)
REAL STAT,DIS1,DIS2,SUMV,R1,R2,AX,AY
DOUBLE PRECISION DSEED
REWIND 1

```

C Fortsetzung Programm 1

```

C
C      EINLESEN VON PMIN,PMAX,DMIN,DMAX,METRIK
C
C      READ (3,300) PMIN,PMAX,DMIN,DMAX,METRIK
      IF (PMIN .LT.5) PMIN=20
      IF ((PMAX.LT.1).OR.(PMAX.GT.15)) PMAX=40
      IF ((DMAX.LT.1).OR.(DMAX.GT.4)) DMAX=5
      IF ((DMIN.LT.1).OR.(DMIN.GT.4)) DMIN=1
      IF ((METRIK.LT.1).OR.(METRIK.GT.99)) METRIK=2
      ITER=1000
      DIS2(I)=0
      DO 10 I=1,800
      R1(I)=0
      R2(I)=0
      DIS1(I)=0
      DIS2(I)=0
10    R(I)=0
      STAT=0
C
C      ERZEUGEN DER UNIFORMVERTEILTEN ZUFALLSZAHLEN ,BERECHNUNG DER
C      DISTANZEN UND DER SCHRITTMASSE ZWISCHEN DEN PERMUTATIONEN.
C
      DSEED=7532.D0
      DO 555 I555=PMIN,PMAX,5
      DO 666 I666=1,DMAX
      DO 11 I=1,800
      DIS1(I)=0
11    DIS2(I)=0
      PKT=I555
      PKT1=PKT-1
      NOD=PKT*(PKT-1)/2
      NPERM=NOD*(NOD-1)/2
      DI=I666
      N=DI*PKT
      N1=DI*PKT1
      DO 21 ITH=1,ITER
      DO 12 IO=1,800
      DIS1(IO)=0
12    DIS2(IO)=0
      CALL ZUF(DSEED,N,R1)
      CALL ZUF(DSEED,N,R2)
C
C
C      HERECHNUNG DER DISTANZEN
C
      K=1
      DO 22 IOIS=1,N1,DI
      J1=IOIS+DI
      DO 23 J=J1,N,DI
      DO 25 IDIM=1,DI
      IR=IOIS+IDIM-1
      JR=J+IDIM-1
      DIS1(K)=DIS1(K)+(R1(IR)-R1(JR))**METRIK
      DIS2(K)=DIS2(K)+(R2(IR)-R2(JR))**METRIK
25    K=1+K
23
22    CONTINUE

```

Fortsetzung : Programm 1

C
C
C
C
C

VERTEILUNG DER DISTANZEN

```

                CALL KENTAU(DIS1,DIS2,NOD,STAT,NPERM)
                ISTAT=IFIX(STAT)
21             R(ISTAT)=R(ISTAT)+1
                PRINT(1,100) DI,PKT,(R(I),I=1,100)
44             DO 45 I=1,5000
45             R(I)=0
                STAT=0
666            CONTINUE
555            CONTINUE
C
100            FORMAT(2X,I2,I3,15I4,/,5(7X,15I4,/,),7X,10I4)
300            FORMAT(5I4)
1000           STOP
                END
                SUBROUTINE KENTAU(AX,AY,NOD,STAT,NPERM)
                DIMENSION AX(NOD),AY(NOD)

```

C
C
C
C
C

SORTIERE X UND Y

```

                NOD1=NOD-1
                DO 1 I=1,NOD1
                I1=I+1
                MIN=AX(I)
                ISTELLE=I
                    DO 2 J=I1,NOD
                    IF (AX(J).GE.MIN) GOTO 2
                    ISTELLE=J
                    MIN=AX(J)
2                CONTINUE
                AX(ISTELLE)=AX(I)
                AX(I)=MIN
                TRANS=AY(ISTELLE)
                AY(ISTELLE)=AY(I)
                AY(I)=TRANS
1            CONTINUE
                STAT=0
                DO 3 I=1,NOD1
                I1=I+1
                DO 4 J=I1,NOD
                IF (AY(I).GT.AY(J)) STAT=STAT+1
4            CONTINUE
3            STAT=STAT*100/NPERM+1
                RETURN
                END
                SUBROUTINE ZUF(DSEED,N,R)
                DIMENSION R(5000)
                DO 1 I=1,N
                R(I)=RANF(DSEED)
1            DSEED=DSEED+23
                RETURN

```

1

Anhang: Programm 2

```

PROGRAM PERM(INPUT,OUTPUT,TAPE1)
*****
C
C
C   PROGRAMM ZUR ERZEUGUNG VON ZUFALLSPERMUTATIONEN DER ZAHLEN
C   1 BIS 780. (780 IST ANZAHL DER DISTANZEN ZWISCHEN 40 PUNKTEN
C   DIE ZAHLEN WERDEN IN FORM VON SYMMETRISCHEN MATRITZEN (ENT-
C   DISTANZMATRITZEN) AUSGEGEBEN. ANZAHL DER PERMUTATIONEN = 10
C
C   *****
C   IMPLICIT INTEGER(A-Z)
C   DOUBLE PRECISION DSEED
C   DIMENSION PER(40,40),ZUF(780)
C
C   ERZEUGUNG DER ZUFALLSZAHLEN
C
C   K=780
C   DSEED=3465723.D0
C   DO 1 I1000=1,1000
C   CALL GGPER(DSEED,K,ZUF)
C       DO 2 J=1,40
C           J1=J+1
C           DO 3 I=1,J
C               IF(I.NE.J) GO TO 4
C               PER(I,J)=0
C               GO TO 3
C           IJ=(J-1)*(J-2)/2+I
C           PER(J,I)=ZUF(IJ)
C           PER(I,J)=(PER(J,I))
C       CONTINUE
C   WRITE (1,100) ((PER(I,J),J=1,40),I=1,40)
100  FORMAT (40(40I3,/),/"*EOR")
C   STOP
C   END

```

Anhang: Programm 3

```
PROGRAM RISK(INPUT,OUTPUT,TAPE1,TAPE2,TAPE3)
```

```
PROGRAMM ZUR BERECHNUNG DER BAYES-RISIKOFUNKTION FUER  
MDS-DARSTELLUNGEN
```

```
*****  
DIMENSION R(5,8,17),ANZ(201),UNR(17),PKT(17),ALPHA(8),RK(8),IKT  
1(17),R1(5,8,17),BETA(201)  
EXTERNAL D01BAZ,FUN1,FUN2,ALOG,EXP  
COMMON SIG1,SIG2  
DATA UNR,RK,PKT,ALPHA/998.,1000.,990.,9*(1000.),3*(224.),194.,131  
2,1.0,0.2,0.1,0.05,0.6,0.5,0.4,0.3,  
34.,5.,6.,7.,8.,9.,10.,11.,12.,13.,14.,15.,20.,25.,30.,35.,40.,  
40,0.01,0.05,0.10,0.15,0.20,0.25,0.30/  
AA=0  
DO 11 I11=1,17  
PP=PKT(I11)  
IKT(I11)=IFIX(PP)  
11 PRINT(3,212)  
212 FORMAT(5X,"DIM PKT",T20,"ERW",T30,"VAR",T40,"P",T50,"U",T60,"SSQ"  
1,/,5X,62("_"))  
BB=1  
NN=64  
DO 1 IK=1,8  
REWIND 1  
SKI=RK(IK)  
DO 2 IP=1,17  
ON=PKT(IP)  
ON3=FLOAT(IP)+3  
ON=NUMBER OF OBJECTS (NUMBER OF POINTS)  
SKI=COEFFICIENT FOR DETERMINING VARIANCE
```

```
HIER BEGINNT DIE BERECHNUNG DER RISIKOFUNKTION
```

```
BERECHNUNG VON SIGMA1 QUADRAT
```

```
OM=ON*(ON-1)/2  
SIG1=(2*OM+5)/18/OM/(OM-1)  
BERECHNUNG VON SIGMA2 QUADRAT  
SIG2=SKI**2
```

```
BERECHNUNG DER GEWICHTE ZUR BESTIMMUNG DES BAYES-RISIKOS
```

```
DO 3 IDI=1,5  
READ (1,100) DIM,PUNKT,(ANZ(I),I=1,201)  
100 FORMAT(T32,F1.0,T39,F2.0,T13,F4.0,/,199(12X,F4.0,/,),12X,F4.0)  
IF ((PUNKT.NE.ON3).OR.(DIM.LT.1).OR.(DIM.GT.5)) PRINT*,"FEHLER 1"
```

Fortsetzung Programm 3

KUMULATIVE HAEUFIGKEITEN UND MITTELWERT

```

C      SUM=0
C      SUMSQ=0
C      ANZV=ANZ(1)
C      DO 10 I10=2,201
C      R10=FLOAT(I10)
C      SUMSQ=SUMSQ+ANZ(I10)*((R10-1)/200)**2
C      SUM=SUM+ANZ(I10)*(R10-1)/200
C      ANZ(I10)=ANZ(I10)+ANZV
10     ANZV=ANZ(I10)
C      ERW=SUM/ONR(IP)
C      ZMOM=SUMSQ/ONR(IP)
C      VAR=ZMOM-(ERW)**2
C      IF(ERW.LE.0) 13,14
13     DO 15 IA=1,8
15     R(IDI,IA,IP)=9.99
C      GO TO 3
14     ABETA=(ERW*ZMOM-ERW**2)/(ERW**2-ZMOM)
C      IF(ABETA.LE.0) ABETA=0
C      BBETA=(ABETA*(1-ERW)/ERW)
C      DO 12 I12=2,201
C      G1=(FLOAT(I12)-1)/200
C      CALL MODBETA(G1,ABETA,BBETA,BET1,IER)
C      IF((IER.EQ.1).OR.(IER.EQ.2)) PRINT*,IER
12     BETA(I12)=BET1
C      BETA(1)=BETA(2)
300    FORMAT(2F3.0,2F12.6)
C
C      QUADRATISCHE ABWEICHUNG
C      QUA=0
C      QUAVO=0
C      DO 16 J=1,201
C      EMP=ANZ(J)/ONR(IP)
C      IF(EMP.LT.0.05) EMP = 1
16     QUA=QUAVO+((EMP-BETA(J))/EMP)**2
C      QUAVO=QUA
C      QUA=QUA/201
C
C      ZAEHLERINTEGRAL 2
C      OMIN=0
C      A2=D01BAF(D01BAZ,OMIN,BB,NN,FUN1,IFAIL)
C      IF ((IFAIL.EQ.1).OR.(IFAIL.EQ.2)) PRINT*,IFAIL
C      NENNERINTEGRAL
C      A3=D01BAF(D01BAZ,AA,BB,NN,FUN2,IFAIL)
C      IF ((IFAIL.EQ.1).OR.(IFAIL.EQ.2)) PRINT*,IFAIL
C      DO 4 IA=1,8
C      AL=ALPHA(IA)
C      AL IST DER ALPHA WERT, DER DEN ANNAHMEREBEREICH DEFINIERT
C      ZAEHLERINTEGRAL 1
C      A1=D01BAF(D01BAZ,AL,BB,NN,FUN1,IFAIL)
C      A4=D01BAF(D01BAZ,AA,AL,NN,FUN2,IFAIL)
C      A5=D01BAF(D01BAZ,AL,BB,NN,FUN2,IFAIL)
C      IF ((IFAIL.EQ.1).OR.(IFAIL.EQ.2)) PRINT*,IFAIL
C      IF(A3.EQ.0) A3=0.000000001
C      R1(IDI,IA,IP)=A1/(A4+A5)
C      RR1=R1(IDI,IA,IP)
C      R2=A2/A3

```

Fortsetzung : Programm 3

```

C
C      BAYESKRISIKO
      AL200=AL*200
      IAL=INT(AL200)
      PDACH=0
      IF (IAL.NE.0) PDACH=BETA(IAL)
      HDACH=BETA(1)
      RDACH=HDACH*RH1+(1-HDACH)*PDACH*R2
      R(IDI,IA,IP)=RDACH
4      CONTINUE
210    FORMAT(6X,F2.0,1X,F2.0,4F10.5,F12.2)
      PRINT(3,210) DIM,ON,ERW,VAR,ABETA,BBETA,QUA
3      CONTINUE
2      CONTINUE
      PRINT(2,200) (SKI,(ALPHA(J),J=1,8),(IKT(IP),(IDI,(R(IDI,IA,IP),
11,8),IDI=1,5),IP=1,17))
200    FORMAT(/,/,/,5X,"KOEFFIZIENT ZUR BESTIMMUNG DER FEHLERVARIANZ K
1F6.4/,/,/,17X,"ALPHA",/,5X,"PKT DIM /",T15,8F6.2,/,5X,
257("-",),/,17(5X,I3.4X,"1",5(8X,2X,I3,T15,8F6.4,/,/))
1      CONTINUE
      STOP
      END

      FUNCTION FAC(OM)
      FAC=1
      M=IFIX(OM)
      IF (M.GE.11) 3,2
3      FAC=0
      RETURN
      DO 1 I=1,M
2      FAC=FAC*I
1      FAC=1/FAC
      RETURN
      END

      FUNCTION FUN1(X)
      COMMON SIG1,SIG2
      E=(((-1)*(X**2*SIG1+(X-0.5)**2)/2/SIG1/SIG2)+ALOG(X))
      IF (E.LT.-675) 2,1
2      FUN1=0
      RETURN
1      FUN1=EXP(E)
      RETURN
      END

      FUNCTION FUN2(X)
      COMMON SIG1,SIG2
      E=(((-1)*(X**2*SIG1+(X-0.5)**2)/2/SIG1/SIG2))
      IF (E.LT.-675) 2,1
2      FUN2=0
      RETURN
1      FUN2=EXP(E)
      RETURN
      END

```

Literaturliste

- (1) Bauer,K. Unveröffentlichter Vortrag im
Rahmen des "workshop MDS" beim
23. Biometrischen Kolloquium
Nürnberg, 1977

- (2) Bauer,K. und Probleme und Perspektiven der
Skarabis,H. Multidimensionalen Skalierung.
Vortrag auf dem 23. Biometri-
schen Kolloquium der deutschen
Region der Internationalen
Biometrischen Gesellschaft,
Nürnberg, 1977

- (3) Borg,I. and A model and algorithm for
Lingoes,J.C. multidimensional scaling with
external constraints on the
distances.
Psychometrica, 1980,45,25-38

- (4) Coombs,C. A Theory of Data.
J.Wiley and Sons,New York,1964

- (5) Coombs,C. Ordmet:A general algorithm for
and McClelland,G. constructing all numerical solutions
to ordered metric structures.
Psychometrika
1975,40,269-290

- (6) Coxon,A.P.M., Multidimensional scaling:
and Jones,C. Exploration to confirmation.
Quality and Quantitiy 1980.
14, 31-73

- (7) Cunningham,J.P. Monotone mapping of simila-
and Shepard,R.N. rities into a general metric

- (15) Guttman, L. Efficacy coefficients for differences among averages. In: Multidimensional Data Representations: When and Why, Mathesis Press, Ann Arbor, 1981, 1-10
- (16) Hotelling, H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. Annals of Mathematical Statistics 1940, 11, 271-283.
- (17) IMSL Mathematical and Statistical Library Reference Manual Edition 8, Houston, Texas 1980
- (18) Kendall, M.G. Rank correlation methods. Charles Griffin and Company, London, 1970
- (19) Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1964, 29, 1-27
- (20) Kruskal, W.H. Ordinal methods of association. Journal of the American Statistical Association, 1958, 53, 814-861
- (21) Konferenz Multidimensional Data Representation: When and Why, im Juni 1980 an der Rheinisch-Westfälischen Technischen Hochschule, Aachen

- (22) Levy, S. and Guttman, L. On the multivariate structure of wellbeing.
Social Indicators Research 1975
2, 361-388
- (23) Levy, S. Use of the mapping sentence for coordinating theory and research: A cross-cultural example
Quality and Quantity, 1976,
10, 117-125
- (24) Levy, S. Lawful roles of facets in social theories. Multidimensional Data Representations: When and Why
Mathesis Press, Ann Arbor, 1981,
65-107
- (25) Lingoes, C.L. and Schönemann, P.H. Alternate measures of fit for the Schönemann-Carroll matrix fitting algorithm.
Psychometrika, 1974, 39, 423-427
- (26) Lingoes, J.C. Indices of configurational similarity.
In: Geometric Representations of Relational Data, Mathesis Press
Ann Arbor 1979, 675-679
- (27) Lingoes, J.C. and Borg, I. An exact significance test for dimensionality and for choosing between rival hypotheses in confirmatory multidimensional scaling.
University of Michigan Computing Technical Report, 1980, 1, 1-22

- (28) Lingo, J.C. Some uses of statistical inference in multidimensional scaling.
In: Multidimensional Data Representations: When and Why, Mathesis Press, Ann Arbor, 1981, 253-279
- (29) Ramsay, J.O. Maximum likelihood estimation in multidimensional scaling. Psychometrika, 1977, 42, 241-265
- (30) Roskam, E.E. The MDS (x) Series of Multidimensional Scaling Programs, Minissa Program
University of Edinburgh Report N° 32, May 1977
- (31) Rothkopf, E.Z. A measure of stimulus similarity and errors in some paired-associate learning tasks. Journal of experimental Psychology, 1957, 53, 94-101
- (32) Sachs, H. Einführung in die Theorie der endlichen Graphen. München, 1971
- (33) Schönemann, P.H. Fitting one matrix to another and Carroll, R.M. under choice of a central dilation and a rigid motion. Psychometrika, 1970, 35, 245-255
- (34) Schubert, G. The dimensions of decisional response: opinion and voting behaviour of the Australian

high court.

In: *Frontiers of Judicial Research*, John Wiley and Sons, New York, 1969, 163-195

- (35) Schulman Ordinal Data: an alternative distribution.
Psychometrika, 1979, 44, 3-20
- (36) Shepard, R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function.
Psychometrika, 1962, 27, 125-139, 219-246
- (37) Shepard, R.N. Representation of structure in similarity data: Problems and Prospects.
Psychometrika, 1974, 39, 373-419
- (38) Shepard, R.N. Presidential address to the psychometric Society.
Psychometrika, 1974, 39, 375
- (39) Skarabis, H. Multidimensional scaling with prescribed structures.
Compstat Lectures 1
Physica-Verlag, Wien, 1978
- (40) Spearman, C. The proof and measurement of association between two things.
American Journal of Psychology, 1904, 15, 88ff
- (41) Spence, I. and Graef, J. The determination of the underlying dimensionality of an

empirically obtained matrix
of proximities.
Multivariate Behavioural Research
1974, 9, 331-341

- (42) Spence, I. A simple approximation for random
rankings stress values.
Mult. Beh. Res. 1979, 14, 355-365
- (43) SPSS Statistical Package for the
Social Sciences
Norman H. Nie u.a.
Mc Graw-Hill, 1975
- (44) BMDP Biomedical Computer Programs
W.J.Dixon, M.B.Brown
University of California Press
Berkeley - Los Angeles - London
1979
- (45) Stegmüller, W. Jenseits von Popper und Car-
nap: Die logischen Grundlagen
des statistischen Schließens.
Springer-Verlag Berlin, 1973
- (46) The MDS (X) Series of Multidimensional Scaling
Programs, Standard Version 3.
Inter-University / Research
Councils Series, Report N^o 51
Januar 1981
- (47) Zinnes, J.L. and Probabilistic multidimensional
Griggs, R.A. unfolding analysis
Psychometrika, 1974, 39, 327-349

Lebenslauf

5. 4. 1951 Geburt in Hannover als erstes von drei Kindern des Diplom-Landwirts Horst Kolleck und der Schneiderin Emmi Kolleck, geb. Balke
- 1957 - 1961 Besuch der evangelischen Grundschulen in Düsseldorf (Eller) und Büttgen bei Neuss
- 1961 - 1969 Besuch des Theodor-Schwann-Gymnasiums in Neuss, neusprachlicher Zweig
9. 6. 1969 Abitur
- 1969 - 1976 Studium der Mathematik, Linguistik und Germanistik an der Technischen Universität Berlin
29. 7. 1976 Diplomprüfung in Mathematik, Abschlußarbeit im Bereich Mathematische Statistik mit dem Thema "Empirische Bayes-Verfahren"
- seit 1976 Mitarbeit in sozialwissenschaftlichen Forschungsvorhaben
- 1978 - 1979 Wissenschaftlicher Mitarbeiter (Bat IIa) im Forschungsprojekt "Integrierte Ausbildungskonzeption Bibliothek, Information, Dokumentation" am Institut für Bibliothekarausbildung der FUB
- seit 1979 Wissenschaftlicher Mitarbeiter mit Teilzeitvertrag (2/3 Bat IIa) am Institut für Soziologie der FU Berlin im Fachgebiet Methodenlehre

Ich bin mit der Sozialpädagogin Susanne Kolleck, geb. Rust verheiratet. Wir haben eine zweijährige Tochter.

